明 細 書

音声認識装置

技術分野

[0001] 本発明は、音声認識装置に関し、詳しくは、話者や、音声認識装置を備えた移動体が移動しても高い精度で音声を認識可能な音声認識装置に関する。

背景技術

[0002] 近年、音声認識技術は、実用化の域に入ってきており、情報の音声入力などに利用され始めている。一方、ロボットの研究開発も盛んとなっており、音声認識技術は、ロボットを実用化するための一つのキー技術ともなっている。すなわち、ロボットと人間との知的なソーシャルインタラクションを行うためには、人間の言葉をロボットが理解する必要があるため、音声認識の精度が重要となっている。

[0003] ところが、実際に人とのコミュニケーションを行うためには、実験室において口元に設置したマイクで音声を入力して行う音声認識とは異なるいくつかの問題がある。

例えば、実際の環境には様々な雑音があり、雑音の中から必要な音声信号を抽出しなければ音声認識をすることができない。また、話者が複数存在する場合にも、同様に認識の対象とする話者の音声のみを抽出する必要がある。また、音声認識においては、一般に隠れマルコフモデル(HMM:Hidden Markov Model)というモデルを利用して内容を特定するが、話者の位置(音響認識装置のマイクを基準とした方向)が異なると、話者の声の聞こえ方も異なることから、認識率に影響を及ぼすという問題がある。

[0004] このようなことから、本発明者を含む研究グループでは、アクティブオーディションにより複数の音源の定位・分離・認識を行う技術を発表している(非特許文献1参照)。

この技術は、人間の耳に相当する位置に2つのマイクを配置し、複数の話者が同時に発話した場合に、一人の発した単語を認識する技術である。詳しくは、2つのマイクから入力された音響信号から、話者の位置を定位し、各話者の音声を分離した上で、音声認識する。この認識の際、移動体(音声認識装置を備えたロボット等)から見て−90° から90° まで10° おきの方向に対する各話者の音響モデルを予め作成して

おく。そして、音声の認識時には、それらの音響モデルを用いて並列に認識プロセスを実行する。

非特許文献1:"A Humanoid Listens to three simultaneous talkers by Integrating Active Audition and Face Recognition" Kazuhiro Nakadai, et al., IJCAI-03 Workshop on Issues in Designing Physical Agents for Dynamic Real-Time Environments: World Modeling, Planning, Learning and Communicating, PP117-124

発明の開示

[0005] しかしながら、前記した従来技術では、話者や移動体が移動する場合には、その都度移動体に対する話者の位置が変化するため、予め用意された音響モデルの方向と異なる方向に話者が位置すると、認識率が低下するという問題があった。

本発明は、このような背景に鑑みてなされたもので、話者や、移動体が移動しても高い精度で認識可能な音声認識装置を提供することを課題とする。

[0006] 前記課題を解決するため、本発明の音声認識装置は、複数のマイクが検出した音響信号から、音声を認識して文字情報に変換する音声認識装置であって、前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、前記複数のマイクが検出した1つ以上の音響信号に基づき、その音響信号に含まれる音声信号の特徴を抽出する特徴抽出部と、断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えるように構成した。

[0007] このような音声認識装置によれば、音源定位部が音源方向を特定し、音響モデル合成部は、音源方向と、方向依存音響モデルとに基づき、その方向に適した音響モデルを合成し、音声認識部がこの音響モデルを使用して音声認識を行う。

[0008] また、前記した音声認識装置においては、音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部を備え、

音源分離部が分離した音声信号に基づき、特徴抽出部が音声信号の特徴を抽出するように構成してもよい。

[0009] このような音声認識装置によれば、音源定位部が音源方向を特定し、音源分離部は、音源定位部が特定した音源方向の音声のみを分離する。そして、音響モデル合成部は、音源方向と、方向依存音響モデルとに基づき、その方向に適した音響モデルを合成し、音声認識部がこの音響モデルを使用して音声認識を行う。

なお、音源分離部が出力する音声信号というのは、音声としての意味を持つ情報であればよく、音声のアナログ信号そのものに限らず、デジタル化、符号化した信号や、周波数分析したスペクトルのデータを含む。

[0010] また、前記した音声認識装置では、前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成することができる。

[0011] また、前記音源定位部は、前記複数のマイクから検出された音響信号の音圧差と位相差を用いて前記特定の話者の音源方向を特定するために、ロボットの頭部などの前記マイクが設けられる部材の表面で散乱する音響信号を音源方向ごとにモデル化した散乱理論を用いることができる。

[0012] さらに、前記した音声認識装置では、前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されるのが好ましい。

[0013] また、前記した音声認識装置では、前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、前記線形和に使用する重みが、学習により決定されるのが好ましい。

[0014] また、前記した音声認識装置では、前記話者を特定する話者同定部をさらに備え、前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、前記音響

モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、前記音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されるのが好ましい。

[0015] また、前記特徴抽出部で抽出された特徴、または前記音源分離部が分離した音声信号について、予め用意した雛形と比較し、前記雛形との違いが予め設定した閾値より大きい領域、例えば周波数領域や、サブバンドを同定し、同定された領域については、その特徴としての信頼性が低いことを示す指標を前記音声認識部へ出力するマスキング部をさらに備えるのが望ましい。

[0016] そして、本発明の他の音声認識装置は、複数のマイクが検出した音響信号から、特定の話者の音声を認識して文字情報に変換する音声認識装置であって、前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、前記音源定位部が特定した音源方向を記憶して前記特定の話者の移動している方向を推定し、その推定された方向から、現在の話者の位置を推定するストリーム追跡部と、前記ストリーム追跡部が推定した現在の話者の位置から定まる音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、前記音源分離部が分離した音声信号の特徴を抽出する特徴抽出部と、断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えて構成することができる。

[0017] このような音声認識装置によれば、任意の方向から発された音響信号の音源方向を特定し、その音源方向に適した音響モデルを使用して音声認識をするので、音声認識率を向上することができる

図面の簡単な説明

[0018] [図1]本発明の実施形態に係る音声認識装置のブロック図である。

[図2]音源定位部の一例を示すブロック図である。

[図3]音源定位部の動作を説明する図である。

[図4]音源定位部の動作を説明する図である。

[図5]聴覚エピポーラ幾何を説明する図である。

[図6]位相差Δφと周波数fの関係を示すグラフである。

[図7]頭部伝達関数の一例を示すグラフである。

[図8]音源分離部の一例を示すブロック図である。

[図9]通過帯域関数の一例を示すグラフである。

[図10]サブバンド選択部の動作を説明する図である。

[図11]通過帯域の一例を図示した平面図である。

[図12](a)および(b)は、ともに特徴抽出部の一例を示すブロック図である。

[図13]音響モデル合成部の一例を示すブロック図である。

[図14]方向依存音響モデルの認識単位とサブモデルを示した図である。

[図15]パラメータ合成部の動作を説明する図である。

[図16](a)および(b)は、ともに重みW_nの一例を示すグラフである。

[図17]重みWの学習方法を説明する図である。

[図18]第2実施形態に係る音声認識装置のブロック図である。

[図19]音響の入力距離差を示す図である。

[図20]第3実施形態に係る音声認識装置のブロック図である。

[図21]ストリーム追跡部のブロック図である。

[図22]音源方向の履歴を図示したグラフである。

発明を実施するための最良の形態

[0019] [第1実施形態]

次に、本発明の実施形態について、適宜図面を参照しながら詳細に説明する。図1は、本発明の実施形態に係る音声認識装置のブロック図である。

図1に示すように、実施形態に係る音声認識装置1は、2つのマイクM_R, M_Lと、マイクM_R, M_Lが検出した音響信号から、話者(音源)の位置を特定する音源定位部10と、音源定位部10が特定した音源方向及び音源定位部10で求めたスペクトルに基づ

いて、特定の方向の音源から来る音響を分離する音源分離部20と、複数の方向についての音響モデルを記憶した音響モデル記憶部49と、音響モデル記憶部49内の音響モデル及び音源定位部10が特定した音源方向に基づいて、その音源方向の音響モデルを合成する音響モデル合成部40と、音源分離部20が分離した特定音源のスペクトルから音響の特徴を抽出する特徴抽出部30と、音響モデル合成部40が合成した音響モデルと、特徴抽出部30が抽出した音響の特徴に基づき音声認識を行う音声認識部50とを備える。これらのうち、音源分離部20は、任意的に用いられる。

本発明では、音響モデル合成部40が生成した、音源の方向に適した音響モデルを利用して音声認識部50が音声認識を行うため、高い認識率が実現される。

[0020] 次に、実施形態に係る音声認識装置1の構成要素であるマイクM_R, M_L、音源定位部10、音源分離部20、特徴抽出部30、音響モデル合成部40、及び音声認識部50についてそれぞれ説明する。

[0021] 《マイクM_R, M_L》

マイクM_R, M_Lは、音を検出して電気信号(音響信号)として出力する一般的なマイクである。本実施形態では、2つとしているが、複数であれば幾つでもよく、例えば3つ、4つを使用しても構わない。マイクM_R, M_Lは、例えば、移動体であるロボットRBの両耳の部分に設けられる。

マイクM_R, M_Lの配置は、音響信号を集音するための一般的な音声認識装置1の正面を決定する。すなわち、マイクM_R, M_Lの集音方向のベクトルの和の方向が音声認識装置1の正面となる。図1に示すように、ロボットRBの頭の左右両脇にマイクM_R, M_Lが1つずつ設けられていれば、ロボットRBの正面が音声認識装置1の正面となる。

[0022] 《音源定位部10》

図2は、音源定位部の一例を示すブロック図であり、図3及び図4は、音源定位部の動作を説明する図である。

音源定位部10は、2つのマイクM_R, M_Lから入力された2つの音響信号から、各話者HMj(図3では、HM1, HM2)の音源方向を定位する。音源定位方法は、マイク

M_R, M_Lに入力された音響信号の位相差を利用する方法、ロボットRBの頭部伝達関数を用いて推定する方法、右と左のマイクM_R, M_Lから入力された信号の相互相関をとる方法などがあり、それぞれ精度を上げるため、種々の改良が加えられているが、ここでは、本発明者が改良した手法を例にして説明する。

[0023] 音源定位部10は、図2に示すように、周波数分析部11、ピーク抽出部12、調波構造抽出部13、IPD計算部14、IID計算部15、聴覚エピポーラ幾何仮説データ16、確信度計算部17、及び確信度統合部18を備える。

これらの各部を、図3及び図4を参照しながら説明する。場面として、ロボットRBに対し、2人の話者HM1, HM2が同時に話しかける場合で説明する。

[0024] 〈周波数分析部11〉

周波数分析部11は、ロボットRBが備える左右のマイクM_R, M_Lが検出した左右の音響信号CR1, CL1から、微小時間Δtの時間長の信号区間を切り出し、左右のチャンネルごとにFFT(高速フーリエ変換)により周波数分析を行う。

例えば、右のマイクM_Rからの音響信号CR1より得られる分析結果がスペクトルCR2であり、左のマイクM_Lからの音響信号CL1より得られる分析結果がスペクトルCL2である。

なお、周波数分析は、バンドパスフィルタなど、他の手法を用いることもできる。

[0025] 〈ピーク抽出部12〉

ピーク抽出部12は、スペクトルCR2, CL2から左右のチャンネルごとに一連のピークを抽出する。ピークの抽出は、スペクトルのローカルピークをそのまま抽出するか、スペクトラルサブトラクション法に基づいた方法(S.F.Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech, Proceedings of 1979 International conference on Acoustics, Speech, and signal Processing (ICASSP-79)参照)で行う。後者の方法は、スペクトルからピークを抽出し、これをスペクトルから減算し、残差スペクトルを生成する。そして、その残差スペクトルからピークが見つからなくなるまでピーク抽出の処理を繰り返す。

前記スペクトルCR2, CL2に対し、ピークの抽出を行うと、例えばピークスペクトルCR3, CL3のようにピークを構成するサブバンドの信号のみが抽出される。

[0026] 〈調波構造抽出部13〉

調波構造抽出部13は、音源が有する調波構造に基づき、左右のチャンネルごとに特定の調波構造を有するピークをグループにする。例えば、人の声であれば、特定の人の声は、基本周波数の音と、基本周波数の倍音とからなるが、人により基本周波数が微妙に異なるので、その周波数の差により、複数の人の声をグループ分けすることができる。調波構造に基づいて同じグループに分けられたピークは、同じ音源から発せられた信号と推定できる。例えば、複数(J人)の話者が同時に話していれば、複数(J個)の調波構造が抽出される。

[0027] 図3においては、ピークスペクトルCR3, CL3の、ピークP1, P3, P5を一つのグループにして調波構造CR41, CL41とし、ピークP2, P4, P6を一つのグループにして調波構造CR42, CL42としている。

[0028] 〈IPD計算部14〉

IPD計算部14は、調波構造抽出部13が抽出した調波構造CR41, CR42, CL41, CL42のスペクトルから、IPD(両耳間位相差)を計算する部分である。

IPD計算部14は、話者HMjに対応する調波構造(例えば、調波構造CR41)に含まれているピーク周波数の集合を$\{f_k \mid k=0...K-1\}$としたとき、各f_kに対応するスペクトルのサブバンドを、右と左の両チャンネル(例えば、調波構造CR41と調波構造CL41)から選択し、次式(1)によりIPD$\Delta\phi(f_k)$を計算する。調波構造CR41と調波構造CL41から計算したIPD$\Delta\phi(f_k)$は、例えば、図4に示す両耳間位相差C51のようになる。ここで、$\Delta\phi(f_k)$は、ある調波構造に含まれるある倍音f_kのIPDであり、Kは、その調波構造に含まれる倍音の数を示す。

[0029] [数1]

$$\Delta\phi(f_k) = \arctan\left(\frac{\Im[S_r(f_k)]}{\Re[S_r(f_k)]}\right) - \arctan\left(\frac{\Im[S_l(f_k)]}{\Re[S_l(f_k)]}\right) \quad \cdots (1)$$

但し、

$\Delta\phi(f_k)$: f_kのＩＰＤ（両耳間位相差）

$\Im[S_r(f_k)]$: 右の入力信号のピーク f_kのスペクトル虚部

$\Re[S_r(f_k)]$: 右の入力信号のピーク f_kのスペクトル実部

$\Im[S_l(f_k)]$: 左の入力信号のピーク f_kのスペクトル虚部

$\Re[S_l(f_k)]$: 左の入力信号のピーク f_kのスペクトル実部

[0030] 〈IID計算部15〉

IID計算部15は、各調波構造にある各倍音について、左のマイクM_Lから入力された音の音圧と、右のマイクM_Rから入力された音の音圧との差(両耳間音圧差)を計算する部分である。

IID計算部15は、話者HMjに対応する調波構造(例えば、調波構造CR41，CL41)に含まれているピーク周波数f_kの倍音に対応するスペクトルのサブバンドを、右と左の両チャンネル(例えば、調波構造CR41と調波構造CL41)から選択し、次式(2)によりIID$\Delta\rho(f_k)$を計算する。調波構造CR41と調波構造CL41から計算したIID$\Delta\rho(f_k)$は、例えば図4に示す両耳間音圧差C61のようになる。

[0031] [数2]

$$\Delta\rho(f_k) = p_r(f_k) - p_l(f_k) \quad \cdots \quad (2)$$

但し、

$\Delta\rho(f_k)$ ： f_kのＩＩＤ（両耳間音圧差）

$p_r(f_k)$ ：右の入力信号のピーク f_kのパワー

$p_l(f_k)$ ：左の入力信号のピーク f_kのパワー

$$p_r(f_k) = 10\log_{10}\left(\Im[S_r(f_k)]^2 + \Re[S_r(f_k)]^2\right)$$

$$p_l(f_k) = 10\log_{10}\left(\Im[S_l(f_k)]^2 + \Re[S_l(f_k)]^2\right)$$

[0032] 〈聴覚エピポーラ幾何仮説データ16〉

聴覚エピポーラ幾何仮説データ16は、図5に示すように、ロボットRBの頭部を想定した球体を上から見たときに、音源Sと、ロボットRBの両耳のマイクM_R，M_Lとの距離差から生じる時間差に基づき想定される位相差のデータである。

聴覚エピポーラ幾何により、位相差$\Delta\phi$は、次式(3)により求められる。ここでは、頭部形状を球と仮定している。

[0033] [数3]

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta) \quad \cdots \quad (3)$$

[0034] ここで、$\Delta\phi$は両耳間位相差(IPD)、vは音速、fは周波数、rは両耳間の距離2rから求まる値、θは音源方向を示す。

式(3)により、各音源方向より発せられた音響信号の周波数fと位相差$\Delta\phi$の関係

は、図6のようになる。

[0035] 〈確信度計算部17〉

確信度計算部17は、IPD及びIIDのそれぞれの確信度を計算する。

—IPD確信度—

IPDの確信度は、話者HMjに対応する調波構造(例えば、調波構造CR41, CL41)が含んでいる倍音f_kがどの方向から来ているらしいかをθの関数として求め、これを確率関数にあてはめる。

まず、f_kのIPDの仮説(予想値)を次式(4)に基づき計算する。

[0036] [数4]

$$\Delta\phi_h(\theta,f_k)=\frac{2\pi f_k}{v}\times r(\theta+\sin\theta) \quad \cdots (4)$$

[0037] $\Delta\phi_h(\theta,f_k)$は、ある調波構造内のk番目の倍音f_kに対して音源方向がθの場合のIPDの仮説(予想値)を示す。IPDの仮説は、例えば音源方向θを、±90°の範囲で5°おきに変化させて計37個の仮説を計算する。もっとも、より細かい角度ごとに計算しても、より大まかな角度ごとに計算してもかまわない。

次に、次式(5)により、$\Delta\phi_h(\theta,f_k)$と$\Delta\phi(f_k)$の差を求め、すべてのピークf_kについて合計する。この差は、仮説と入力との距離を表し、θが話者のいる方向に近いと小さく、遠いと大きくなる。

[0038] [数5]

$$d(\theta)=\frac{1}{K}\sum_{k=0}^{K-1}\frac{\left(\Delta\phi_h(\theta,f_k)-\Delta\phi(f_k)\right)^2}{f_k} \quad \cdots (5)$$

[0039] 得られた$d(\theta)$を、次式(6)の確率密度関数に代入し、確信度$B_{IPD}(\theta)$を得る。

[0040] [数6]

$$B_{IPD}(\theta)=\int_{-\infty}^{X(\theta)}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx \quad \cdots (6)$$

ここで、$X(\theta)=(d(\theta)-m)/(\sqrt{(s/n)})$、mは、$d(\theta)$の平均、sは$d(\theta)$の分散であり、nはIPDの仮説の個数(本実施形態では37個)である。

[0041] —IID確信度—

IIDの確信度は、以下のようにして求める。まず、話者HMjに対応ずる調波構造が含む倍音の音圧差の合計を次式(7)で計算して求める。

[0042] [数7]

$$S = \sum_{k=0}^{K-1} \Delta \rho (f_k) \quad \cdots (7)$$

[0043] ここで、Kは、その調波構造に含まれる倍音の数を示し、$\Delta \rho (f_k)$は、IID計算部15で求めたIIDである。

次に、表1を利用して、音源方向の右らしさ、正面らしさ、左らしさを確信度とする。なお、表1は、実験的に得られた値である。

例えば、表1を参照して、仮説の音源方位θが40° で、音圧差Sが正であれば確信度$B_{IID}(\theta)$は、左上の欄を参照して0．35とする。

[0044] [表1]

θ		90° ～30°	30° ～－30°	－30° ～－90°
S	＋	0．35	0．5	0．65
	－	0．65	0．5	0．35

[0045] 〈確信度統合部18〉

確信度統合部18は、Dempster-Shafer理論に基づき、IPDとIIDの確信度$B_{IPD}(\theta)$、$B_{IID}(\theta)$を次式(8)によって統合し、統合確信度$B_{IPD+IID}(\theta)$を計算する。そして、統合確信度$B_{IPD+IID}(\theta)$が最も大きくなる音源方向θを、話者HMjのいる方向とし、以下θ_{HMj}とする。

[0046] [数8]

$$B_{IPD+IID}(\theta) = 1 - (1 - B_{IPD}(\theta))(1 - B_{IID}(\theta)) \quad \cdots (8)$$

[0047] 以上のような聴覚エピポーラ幾何を使用した仮説に代えて、頭部伝達関数を用いた仮説データ、又は散乱理論に基づく仮説データを用いることもできる。

(頭部伝達関数仮説データ)

頭部伝達関数仮説データは、ロボット周囲から発せられたインパルスより得られる、マイクM_RとマイクM_Lで検出した音の位相差及び音圧差である。

頭部伝達関数仮説データは、－90° から90° の間の適当な間隔(例えば5°)の

方向から発したインパルスを、マイクM_R, M_Lで検出し、それぞれを周波数分析して周波数fに対する位相応答及び振幅応答を求め、その差を計算することによって得られる。

得られた頭部伝達関数仮説データは、図7(a)のIPD及び(b)のIIDのようになる。

頭部伝達関数を用いる場合には、IPDだけではなく、IIDについてもある音源方向から来た音の周波数とIIDの関係が求められるので、IPDとIIDの両方について距離データd(θ)を作ってから確信度を求める。仮説データの作成方法は、IPDとIIDで変わりはない。

聴覚エピポーラ幾何を利用した仮説データの作成方法と異なり、計算ではなく計測で、各音源方向で発せられた信号に対する周波数fとIPDの関係を求める。すなわち、図7(a), (b)にある実測値から、それぞれの仮説と入力との距離であるd(θ)を直接計算する。

[0048] (散乱理論に基づく仮説データ)

散乱理論は、音を散乱する物体、例えばロボットの頭部による散乱波を考慮して、IPD、IIDの双方を計算的に推定する理論である。ここでは、音を散乱する物体の内、マイクの入力に主に影響を与える物体はロボットの頭部であると仮定し、これを半径aの球と仮定する。また頭部の中心の座標を極座標の原点とする。

点音源の位置をr_0、観測点をrとすると、観測点における直接音によるポテンシャルは、次式(9)によって定義される。

[数9]

$$V^i = \frac{v}{2\pi R\, f} e^{i\frac{2\pi R f}{v}} \quad \cdots (9)$$

但し、

f:点音源の周波数

v:音速

R:点音源と観測点の距離

また、観測点rを頭部表面とすると、直接音と散乱音によるポテンシャルは、

J.J.Bowman, T.B.A. Senior, and P.L.E. Uslenghi: Electromagnetic and Acoustic

Scattering by Simple Shapes. Hemisphere Publishing Co., 1987.などに開示されているように、次式(10)で定義される。

[数10]

$$S(\theta,\mathrm{f}) = V^i + V^S$$

$$= -\left(\frac{v}{2\pi a\,\mathrm{f}}\right)^2 \sum_{n=0}^{\infty}(2n+1)P_n(\cos\theta)\frac{h_n^{(1)}\left(\frac{2\pi r_0}{v}\mathrm{f}\right)}{h_n^{(1)'}\left(\frac{2\pi a}{v}\mathrm{f}\right)} \quad \cdots (10)$$

但し、

V^S:散乱音によるポテンシャル

P_n:第一種Legendre関数

$h_n^{(1)}$:第一種球ハンケル関数

M_Rの極座標を(a, π／2, 0)、M_Lの極座標を(a, −π／2, 0)とすると、それぞれにおけるポテンシャルは、次式(11)、(12)で表される。

[数11]

$$S_L(\theta,\mathrm{f}) = S(\frac{\pi}{2}-\theta,\mathrm{f}) \quad \cdots (11)$$

[数12]

$$S_R(\theta,\mathrm{f}) = S(-\frac{\pi}{2}-\theta,\mathrm{f}) \quad \cdots (12)$$

従って、散乱理論に基づく位相差IPD$\Delta\phi_S(\theta,\mathrm{f})$と音圧差IID$\Delta\rho_S(\theta,\mathrm{f})$は、それぞれ次式(13)、(14)により求められる。

[数13]

$$\Delta\phi_S(\theta,\mathrm{f}) = \arg(S_L(\theta,\mathrm{f})) - \arg(S_R(\theta,\mathrm{f})) \quad \cdots (13)$$

[数14]

$$\Delta\rho_S(\theta,\mathrm{f}) = 20\log_{10}\frac{|S_L(\theta,\mathrm{f})|}{|S_R(\theta,\mathrm{f})|} \quad \cdots (14)$$

[0049] そして、前記(4)式の$\Delta\phi_h(\theta,f_k)$を前記(13)式のIPD$\Delta\phi_S(\theta,\mathrm{f})$に置き換え、前

記した聴覚エピポーラ幾何を用いた場合と同じ手順で$B_{IPD}(\theta)$を求める。

すなわち、$\Delta\phi_s(\theta, f_k)$と$\Delta\phi(f_k)$の差を求め、すべてのピーク$f_k$について合計して$d(\theta)$を求め、得られた$d(\theta)$を、前記式(6)の確率密度関数に代入し、確信度$B_{IPD}(\theta)$を得る。

[0050] IIDもIPDと同じ方法で$d(\theta)$と$B_{IID}(\theta)$を計算する。具体的には、$\Delta\phi$を$\Delta\rho$とし、前記(4)式の$\Delta\phi_h(\theta, f_k)$を前記(14)式のIPD$\Delta\rho_S(\theta, f_k)$でに置き換える。そして、$\Delta\rho_S(\theta, f_k)$と$\Delta\rho(f_k)$の差を求め、すべてのピーク$f_k$について合計して$d(\theta)$を求め、得られた$d(\theta)$を、前記式(6)の確率密度関数に代入し、確信度$B_{IID}(\theta)$を得る。

[0051] このように散乱理論に基づいて音源方向を推定すると、ロボットの頭部の表面に沿って散乱する音声、例えば後頭部を回り込む音の影響を考慮して、音源方向と位相差、および音源方向と音圧差の関係をモデル化できるので、音源方向の推定精度が向上する。特に、音源が側方にある場合は、後頭部を回り込んで音源と反対方向にあるマイクに到達する音のパワーは比較的大きいため、散乱理論を用いることによって音源方向の推定精度が向上する。

[0052] 《音源分離部20》

音源分離部20は、音源定位部10により定位された各音源方向の情報、並びに音源定位部で計算したスペクトル(例えばスペクトルCR2)により、各話者HMjの音響(音声)信号を分離する部分である。音源分離方法には、ビームフォーミング、ナルフォーミング、ピーク追跡、指向性マイク、ICA(Independent Component Analysis:独立成分分析)など、従来からある手法を用いることができるが、ここでは、本発明者が開発したアクティブ方向通過型フィルタによる方法について説明する。

音源方向の情報を利用して音源を分離する場合、音源の方向がロボットRBの正面から離れるにつれ、2本のマイクを用いて推定した音源方向情報の精度を期待できなくなる。そこで、本実施形態では、正面方向の音源については通過させる方向の範囲を狭く、正面から離れた音源では広くとるように通過帯域をアクティブに制御して、音源の分離精度を向上させる。

[0053] 具体的には、音源分離部20は、図8に示すように、通過帯域関数21と、サブバンド選択部22とを有する。

[0054] 〈通過帯域関数21〉

通過帯域関数21は、図9に示したように、音源方向と通過帯域幅の関数で、音源方向が、正面(0°)から離れるにつれ、方向情報の精度を期待できなくなることから、音源方向が正面から離れるほど通過帯域幅が大きくなるように予め設定した関数である。

[0055] 〈サブバンド選択部22〉

サブバンド選択部22は、スペクトルCR2, CL2の各周波数の値(これを「サブバンド」という)から、特定の方向から来たと推測されるサブバンドを選択する部分である。

サブバンド選択部22では、図10に示すように、音源定位部10で生成した左右の入力音のスペクトルCR2, CL2から、各スペクトルのサブバンドについて、前記式(1)、(2)に従い、IPD $\Delta\phi(f_i)$及びIID $\Delta\rho(f_i)$を計算する(図10の両耳間位相差C52, 両耳間音圧差C62参照)。

そして、音源定位部10で得られた θ_{HMj} を抽出すべき音源方向とし、通過帯域関数21を参照して、θ_{HMj} に対応する通過帯域幅 $\delta(\theta_{HMj})$ を取得する。取得した通過帯域幅 $\delta(\theta_{HMj})$ を用いて、通過帯域の最大値 θ_h と最小値 θ_l を次式(15)により求める。通過帯域Bは、方向として平面図で図示すると、例えば図11のようになる。

[0056] [数15]

$$\left.\begin{aligned}\theta_l &= \theta_{HMj} - \delta(\theta_{HMj}) \\ \theta_h &= \theta_{HMj} + \delta(\theta_{HMj})\end{aligned}\right\} \quad \cdots (15)$$

[0057] 次に、θ_l と θ_h に対応するIPDとIIDを推定する。これらの推定には、予め計測、又は計算した伝達関数を利用する。伝達関数は、音源方向 θ から来る信号に対して周波数fとIPD、IIDをそれぞれ関係づけている関数で、前記したエピポーラ幾何や、頭部伝達関数、散乱理論などを用いる。推定したIPDは、例えば図10の両耳間位相差C53における $\Delta\phi_l(f)$, $\Delta\phi_h(f)$ であり、推定したIIDは、例えば図10の両耳間音圧差C63における $\Delta\rho_l(f)$, $\Delta\rho_h(f)$ である。

[0058] 次に、音源方向 θ_{HMj} に対して、ロボットRBの伝達関数を利用して、スペクトルCR2またはCL2の周波数 f_i に応じ、周波数 f_i が所定の閾値周波数 f_{th} より小さければIPDによりサブバンドを選択し、大きければIIDによりサブバンドを選択する。すなわち、以下

の条件式(16)を満たすサブバンドを選択する。

[0059] [数16]

$$\left.\begin{array}{l} f_i < f_{th} : \Delta\phi_l(f_i) \leqq \Delta\phi(f_i) \leqq \Delta\phi_h(f_i) \\ f_i \geqq f_{th} : \Delta\rho_l(f_i) \leqq \Delta\rho(f_i) \leqq \Delta\rho_h(f_i) \end{array}\right\} \quad \cdots (16)$$

[0060] ここで、f_{th}は、フィルタリングの判断基準にIPDとIIDのどちらを用いるかを決める閾値周波数である。

この条件式によれば、例えば、図10の両耳間位相差C53においては、周波数f_{th}より低い周波数で、IPDが$\Delta\phi_l(f)$と$\Delta\phi_h(f)$の間にある周波数f_iのサブバンド(斜線部)が選択される。一方、図10の両耳間音圧差C63においては、周波数f_{th}より高い周波数で、IIDが$\Delta\rho_l(f)$と$\Delta\rho_h(f)$の間にあるサブバンド(斜線部)が選択される。この選択されたサブバンドからなるスペクトルを本明細書において「選択スペクトル」という。

[0061] 以上、本実施形態の音源分離部20について説明したが、音源分離の方法には、この他に指向性マイクを利用した方法がある。即ち、指向性が狭いマイクをロボットRBに設けておき、音源定位部10で得られた音源方向θ_{HMj}の方向に指向性マイクを向けるよう、顔の向きを変えれば、その方向から来る音声だけを取得することができる。

この指向性マイクによる方法の場合、1つの指向性マイクしかない場合には、1人の音声しか取得できないという問題もあるが、複数の指向性マイクを所定角度おきに設けておき、音源方向の指向性マイクからの音声信号を利用するようにすれば、複数人の音声の同時取得も可能である。

[0062] 《特徴抽出部30》

特徴抽出部30は、音源分離部20で分離された音声スペクトルあるいは分離をしないスペクトルCR2(またはCL2)(以下、音声認識に使用する場合に「認識用スペクトル」という)から音声認識に必要な特徴を抽出する部分である。音声の特徴としては、音声を周波数分析した線形スペクトルや、メル周波数スペクトル、メル周波数ケプストラム係数(MFCC:Mel-Frequency Cepstrum Coefficient)を用いることができる。本実施形態では、MFCCを用いる場合で説明する。なお、線形スペクトルを特徴として用いる場合は、特徴抽出部30は、特に処理を行わない。また、メル周波数スペクトルを

用いる場合は、コサイン変換(後述)を行わない。

[0063] 特徴抽出部30は、図12(a)に示すように、対数変換部31、メル周波数変換部32、及びコサイン変換部33を有する。

対数変換部31は、サブバンド選択部22(図8参照)が選択した認識用スペクトルの振幅を対数に変換して、対数スペクトルを得る。

メル周波数変換部32は、対数変換部31が生成した対数スペクトルを、メル周波数のバンドパスフィルタに通し、周波数がメルスケールに変換されたメル周波数対数スペクトルを得る。

コサイン変換部33は、メル周波数変換部32が生成したメル周波数対数スペクトルをコサイン変換する。このコサイン変換により得られた係数がMFCCとなる。

[0064] また、雑音などによって入力音声が変形している場合は、そのスペクトルサブバンドを特徴として信用しないよう、図12(b)に示すように指標(0から1)を付与するマスキング部34を、特徴抽出部30の中または後に任意的に追加してもよい。

図12(b)の例について具体的に説明すると、特徴抽出部30が任意的にマスキング部34を含む場合、単語辞書59は、単語に対応してその単語の時系列スペクトルを有する。ここでは、この時系列スペクトルを「単語音声スペクトル」とする。

単語音声スペクトルは、雑音がない環境下で単語を発声した音声を周波数分析して得られる。特徴抽出部30に認識用スペクトルが入力されると、入力音声に含まれていると推測された単語の単語音声スペクトルが想定音声スペクトルとして単語辞書から選別される。ここでは、認識用スペクトルと時間長が最も近いものを想定音声スペクトルとして推測する。認識用スペクトルと想定音声スペクトルは、それぞれ対数変換部31、メル周波数変換部32、コサイン変換部33を経てMFCCに変換される。以下、認識用スペクトルのMFCCを「認識用MFCC」、想定音声スペクトルのMFCCを「想定MFCC」とする。

マスキング部34は、認識用MFCCと想定MFCCの差を求め、予め想定した閾値より大きい場合は0を、小さい場合は1を、MFCCの特徴量ベクトルの各特徴ごとに付与する。これを指標ωとして認識用MFCCと合わせて音声認識部50に出力する。

想定音声スペクトルを選別する際、1つだけではなく、複数選別してもよい。また、選

別せずに全ての単語音声スペクトルを用いてもよい。その場合には、すべての想定音声スペクトルについて指標ωを求め、音声認識部50に出力する。

[0065] なお、指向性マイクを用いて音源分離を行う場合には、指向性マイクから得られた分離音声に対し、FFTやバンドパスフィルタなどの一般的な周波数分析手法を用いてスペクトルを得る。

[0066] 《音響モデル合成部40》

音響モデル合成部40は、音響モデル記憶部49に記憶された方向依存音響モデルから、定位された各音源方位に応じた音響モデルを合成する部分である。

音響モデル合成部40は、図13に示すように、コサイン逆変換部41、線形変換部42、指数変換部43、パラメータ合成部44、対数変換部45、メル周波数変換部46、及びコサイン変換部47を有し、音響モデル記憶部49に記憶された方向依存音響モデルH(θ_n)を参照してθ方向の音響モデルを合成する。

[0067] 〈音響モデル記憶部49〉

音響モデル記憶部49には、ロボットRBの正面を基準とした方向θ_nごとに、方向θ_nに適した音響モデルである方向依存音響モデルH(θ_n)が記憶されている。方向依存音響モデルH(θ_n)は、特定の方向θ_nから発せられた人物の音声の特徴を、隠れマルコフモデル(HMM)で学習させたものである。各方向依存音響モデルH(θ_n)は、図14に示すように、例えば音素を認識単位とし、音素ごとに対応するサブモデルh(m, θ_n)を記憶している。なお、サブモデルは、モノフォン、PTM、バイフォン、トライフォンなど他の認識単位で作成してもよい。

サブモデルh(m, θ_n)の数は、例えば方向θ_nについて−90° ー90° まで30° おきに7個のモデルを持ち、サブモデルを40個のモノフォンで構成しているとすれば、合計7×40＝280個となる。

サブモデルh(m, θ_n)は、状態数、各状態の確率密度分布、状態遷移確率の各パラメータを有している。本実施形態では、各音素の状態数は、前部(状態1)、中間部(状態2)、後部(状態3)の3つに固定している。また、本実施形態では、確率密度分布は、正規分布に固定するが、確率密度分布は、正規分布または他の分布の1つ以上の混合分布であってもよい。したがって、本実施形態では、状態遷移確率Pと、正

規分布のパラメータ、つまり平均μ及び標準偏差σを学習させる。

[0068] サブモデルh(m, θ_n)の学習データは次のようにして作成する。

ロボットRBに対し、音響モデルを作成したい方向から、特定の音素からなる音声信号を図示しないスピーカにより発する。そして、検出した音響信号を特徴抽出部30によりMFCCに変換し、後述する音声認識部50で音声認識させる。すると、認識した音声が、音素ごとにどのくらいの確率であるかが結果として得られるが、この結果に対し、特定の方向の特定の音素であるという教師信号を与えることで音響モデルを適応学習させる。そして、サブモデルを学習するのに十分な種類(例えば、異なる話者)の音素や単語を学習させる。

なお、学習用音声を発する際、音響モデルを作成したい方向とは異なる方向から、別の音声をノイズとして発してもよい。この場合は、前記した音源分離部20により音響モデルを作成したい方向の音響のみを分離した上で、特徴抽出部30によりMFCCに変換する。また、これらの学習は、音響モデルを不特定話者のモデルとして持たせたい場合には、不特定の話者の声で学習させればよいし、特定話者ごとにモデルを持たせたい場合には、特定話者ごとに学習させればよい。

[0069] コサイン逆変換部41から指数変換部43は、確率密度分布のMFCCを線形スペクトルに戻す。つまり、確率密度分布について、特徴抽出部30と逆の操作をする。

[0070] 〈コサイン逆変換部41〉

コサイン逆変換部41は、音響モデル記憶部49が記憶している方向依存音響モデルH(θ_n)が有するMFCCについてコサイン逆変換してメル対数スペクトルを生成する。

[0071] 〈線形変換部42〉

線形変換部42は、コサイン逆変換部41により生成されたメル対数スペクトルの周波数を線形周波数に変換し、対数スペクトルを生成する。

[0072] 〈指数変換部43〉

指数変換部43は、線形変換部42により生成された対数スペクトルの強度を指数変換し、線形スペクトルを生成する。線形スペクトルは、平均μ、標準偏差σの確率密度分布として得られる。

[0073] 〈パラメータ合成部44〉

パラメータ合成部44は、図15に示すように、方向依存音響モデルH(θ_n)にそれぞれ重みをかけた上でそれらの和をとり、音源方向θ_{HMj}の音響モデルH(θ_{HMj})を合成する。方向依存音響モデルH(θ_n)にある各サブモデルは、それぞれコサイン逆変換部41から指数変換部43により、線形スペクトルの確率密度分布に変換され、それぞれ、平均μ_{1nm}, μ_{2nm}, μ_{3nm}, 標準偏差σ_{1nm}, σ_{2nm}, σ_{3nm}, 状態遷移確率P_{11nm}, P_{12nm}, P_{22nm}, P_{23nm}, P_{33nm}のパラメータを持っている。そして、これらのパラメータを、予め学習によって求められ、音響モデル記憶部49に記憶されている重みと内積して、音源方向θ_{HMj}の音響モデルを合成する。つまり、パラメータ合成部44は、方向依存音響モデルH(θ_n)の線形和により音源方向θ_{HMj}の音響モデルを合成している。なお、重み$W_{n\theta HMj}$の設定の仕方は後述する。

[0074] H(θ_{HMj})にあるサブモデルを合成する場合には、状態1の平均$\mu_{1\theta HMjm}$を次式(17)により求める。

[0075] [数17]

$$\mu_{1\theta HMjm} = \frac{1}{\sum_{n=1}^{N} W_{n\theta HMj}} \sum_{n=1}^{N} W_{n\theta HMj}\mu_{1nm} \quad \cdots (17)$$

[0076] 平均$\mu_{2\theta HMjm}$, $\mu_{3\theta HMjm}$についても同様にして求めることができる。

[0077] また、状態1の標準偏差$\sigma_{1\theta HMjm}$の合成については、共分散$\sigma_{1\theta HMjm}^2$を次式(18)により求める。

[数18]

$$\sigma_{1\theta HMjm}^2 = \frac{1}{\sum_{n=1}^{N} W_{n\theta HMj}} \sum_{n=1}^{N} W_{n\theta HMj}\sigma_{1nm}^2 \quad \cdots (18)$$

[0078] 標準偏差$\sigma_{2\theta HMjm}$, $\sigma_{3\theta HMjm}$についても同様にして求めることができる。
得られたμとσにより、確率密度分布を求めることができる。

[0079] また、状態1の状態遷移確率$P_{11\theta HMjm}$の合成については、次式(19)により求める。

[0080] [数19]

$$P_{11\theta HMjm}=\frac{1}{\sum_{n=1}^{N}W_{n\theta HMj}}\sum_{n=1}^{N}W_{n\theta HMj}P_{11nm} \quad \cdots (19)$$

[0081] 状態遷移確率$P_{12\theta HMjm}$, $P_{22\theta HMjm}$, $P_{23\theta HMjm}$, $P_{33\theta HMjm}$についても同様にして求めることができる。

[0082] 次に、対数変換部45からコサイン変換部47により、確率密度分布を線形スペクトルからMFCCに変換し直す。すなわち、対数変換部45は、対数変換部31と、メル周波数変換部46は、メル周波数変換部32と、コサイン変換部47は、コサイン変換部33と同様であるので、詳細な説明を省略する。

[0083] なお、単一正規分布ではなく、混合正規分布の形で合成する場合には、前記した平均μ、標準偏差σの計算に代えて次式(20)により確率密度分布$f_{1\theta HMjm}(x)$を求める。

[0084] [数20]

$$f_{1\theta HMjm}(x)=\frac{1}{\sum_{n=1}^{N}W_{n\theta HMj}}\sum_{n=1}^{N}W_{n\theta HMj}f_{1nm}(x) \quad \cdots (20)$$

[0085] 確率密度分布$f_{2\theta HMjm}(x)$, $f_{3\theta HMjm}(x)$についても同様にして求めることができる。

[0086] パラメータ合成部44は、このようにして得られた音響モデルを、音響モデル記憶部49に記憶させる。

なお、このような音響モデルの合成は、音声認識装置1が作動している間、パラメータ合成部44がリアルタイムに行う。

[0087] 〈重み$W_{n\theta HMj}$の設定〉

重み$W_{n\theta HMj}$は、音源方向θ_{HMj}に対応する音響モデルを合成するときに、各方向依存音響モデル$H(\theta_n)$に対して設定するもので、$H(\theta_n)$に含まれるすべてのサブモデル$h(m, \theta_n)$に対して用いる重み$W_{n\theta HMj}$を設定してもよいし、あるいは各サブモデル$h(m, \theta_n)$に対応する重み$W_{mn\theta HMj}$を設定してもよい。基本的には、音源が正面にある場合の重み$W_{n\theta 0}$を定める関数$f(\theta)$をあらかじめ設定しておき、音源方向θ_{HMj}に対応する音響モデルを合成する際に、$f(\theta)$をθ軸方向にθ_{HMj}移動($\theta \rightarrow \theta-\theta_{HMj}$とする)した関数$f(\theta)$を求め、これを参照して$W_{n\theta HMj}$を設定する。

[0088] 〈関数f(θ)の作成〉

[A] f(θ)を経験的に求める方法

f(θ)を経験的に求める場合は、経験的に得られた定数aを用いて次式のように表す。

$f(\theta) = a\theta + \alpha$ ($\theta < 0$, $\theta = -90°$ のとき$f(\theta) = 0$)

$f(\theta) = -a\theta + \alpha$ ($\theta \geqq 0$, $\theta = 90°$ のとき$f(\theta) = 0$)

ここで、定数a＝1．0とすれば、音源が正面にある場合のf(θ)は、図16(a)のようになる。また、f(θ)をθ軸方向にθ_{HMj}移動したのが図16(b)である。

[0089] [B] f(θ)を学習によって求める方法

f(θ)を学習によって求める場合は、例えば次のような学習をする。

音源が正面にあるときの任意の音素mの重みを$W_{mn\theta 0}$とする。最初に適当な初期値の重みの値の$W_{mn\theta 0}$を設定しておき、この$W_{mn\theta 0}$を用いて合成した音響モデルH(θ_0)でmを含む適当な音素列、例えば音素列[m m′ m″]を認識させる試行を行う。具体的には、正面に設置したスピーカから、前記音素列を発し、これを認識させる。ここで、学習データは、1つの音素m自体であってもよいのであるが、音素が複数つながった音素列で学習させた方が良い学習結果が得られるため、音素列を使用している。

この時の認識結果が、例えば図17である。図17では、初期値の$W_{mn\theta 0}$を用いて合成した音響モデルH(θ_0)での認識結果が1行目であり、2行目以下のH(θ_n)が方向θ_nの方向依存音響モデルH(θ_n)を使用したときの認識結果である。例えば、音響モデルH(θ_{90})での認識結果は音素列[/x//y//z/]であり、音響モデルH(θ_0)での認識結果は、音素列[/x//y/m″]であったことを示す。

1回目の試行後、まず1音素目を見て、図17の正面から$\theta = \pm 90°$ の範囲に一致する音素が認識された場合、その方向に対応するモデルの重み$W_{mn\theta 90}$をΔd増加させる。Δdは実験的に求め、例えば0．05とする。そして、一致する音素が認識されない場合、その方向に対応するモデルの重み$W_{mn\theta 0}$をΔd／(n−k)減少させる。つまり、正解を出した方向依存音響モデルの重みは大きくし、正解を出さなかった方向依存音響モデルの重みは減少させる。

[0090] 例えば、図17の場合では、H(θ_n)とH(θ_{90})が一致しているので、対応する重み$W_{mn\theta}$と重み$W_{m90\theta 0}$をΔd増加させ、それ以外の重みを2Δd／(n−2)減少させる。

一方、1音素目に一致する音素を認識した方向θ_nが無い場合、他の方向に対して重みの大きい、優勢な方向依存音響モデルH(θ_n)があれば、その方向依存音響モデルH(θ_n)の重みをΔd減少させ、それ以外のモデルの重みをkΔd／(n−k)増加させる。つまり、どの方向依存音響モデルH(θ_n)も認識できなかったということは、現在の重みの分配が良くない可能性があるから、現在の重みが優勢な方向について重みを減少させる。

優勢であるかどうかは、重みが予め定められた閾値(ここでは0．8とする)より大きいかどうかで判断する。優勢な方向依存音響モデルH(θ_n)がなければ、最大の重みのみをΔd減少させ、その他の方向依存音響モデルH(θ_n)の重みをΔd／(n−1)増加させる。

そして、更新された重みを用いて、前記した試行を繰り返す。

そして、音響モデルH(θ_{90})の認識結果が、正解mとなったときに、繰り返しを終了し、次の音素m′の認識および学習へ移るか、または学習を終了する。学習を終了した場合、ここで得られた重み$W_{mn\theta 90}$がf(θ)となる。次の音素m′へ移る場合は、すべての音素について学習し、得られた$W_{mn\theta 90}$を平均したものがf(θ)となる。

これを平均せず、各サブモデルh(m，θ_n)に対応する重み$W_{mn\theta HMj}$をf(θ)にしてもよい。

なお、所定の回数(例えば0．5／Δd回)繰り返しても、音響モデルH(θ_{HMj})の認識結果が正解に至らない場合、例えばmの認識がうまくいかなかった場合には、次の音素m′の学習へ移り、最終的にうまく認識できた音素(例えばm′)の重みの分布と同じ値で重みを更新する。

また、音響モデルを合成するたびにf($\theta-\theta_{HMj}$)を求めるのではなく、予め適当なθ_{HMj}について、H(θ_n)に含まれるすべてのサブモデルh(m，θ_n)(表2参照)が用いる重み$W_{n\theta HMj}$または各サブモデルh(m，θ_n)に対応する$W_{n\theta HMj}$を求めた表3を作成しておいてもよい。なお、表2および表3において、添え字の1・・・m・・・Mは音素を表し、1・・・n・・・Nは方向を表す。

[表2]

$H(\theta_1)$	$H(\theta_2)$	…	$H(\theta_n)$	…	$H(\theta_N)$
$h(1,\theta_1)$	$h(1,\theta_2)$	…	$h(1,\theta_n)$	…	$h(1,\theta_N)$
⋮	⋮	…	⋮	…	⋮
$h(m,\theta_1)$	$h(m,\theta_1)$	…	$h(m,\theta_n)$	…	$h(m,\theta_N)$
⋮	⋮	…	⋮	…	⋮
$h(M,\theta_1)$	$h(M,\theta_2)$	…	$h(M,\theta_n)$	…	$h(M,\theta_N)$

[表3]

W_1	W_2	…	W_n	…	W_N
w_{11}	w_{12}	…	w_{1n}	…	w_{1N}
⋮	⋮	…	⋮	…	⋮
w_{m1}	w_{m2}	…	w_{mn}	…	w_{mN}
⋮	⋮	…	⋮	…	⋮
w_{M1}	w_{M2}	…	w_{Mn}	…	w_{MN}

[0091] このようにして学習して得られた重みは、音響モデル記憶部49に記憶させる。

[0092] 《音声認識部50》

音声認識部50は、音源方向θ_{HMj}に対応して合成された音響モデル$H(\theta_{HMj})$を用いて、分離された各話者HMjの音声あるいは入力音声から抽出した特徴を認識して文字情報とし、単語辞書59を参照して言葉を認識し、認識結果を出力する。この音声認識の方法は一般的な隠れマルコフモデルを利用した認識方法なので、詳細な説明は省略する。

なお、マスキング部を特徴抽出部30の中または後に設けて、MFCCの各サブバンドの信用度を示す指標ωが付与されている場合には、音声認識部50は、入力された

特徴に次式(21)のような処理を行ってから認識する。

[数21]

$$\left.\begin{aligned} x_r &= 1 - x_n \\ x_n(i) &= x(i) \times \omega(i) \end{aligned}\right\} \quad \cdots (16)$$

x_r ：音声認識に用いる特徴
x ：MFCC
i ：MFCCの成分
x_n ：xのうち信用できない成分

そして、得られた出力確率と状態遷移確率を用いて、一般的な隠れマルコフモデルを利用した認識方法と同様に認識を行う。

[0093] 以上のように構成された、音声認識装置1による動作を説明する。

図1に示すように、ロボットRBのマイクM_R, M_Lに、複数の話者HMj(図3参照)の音声が入力される。

そして、マイクM_R, M_Lが検出した音響信号の音源方向が音源定位部10で定位される。音源定位は、前記したように周波数分析、ピーク抽出、調波構造の抽出、IPD・IIDの計算の後、聴覚エピポーラ幾何に基づいた仮説データを利用して確信度を計算する。そして、IPDとIIDの確信度を統合して最も可能性が高いθ_{HMj}を音源方向とする(図2参照)。

[0094] 次に、音源分離部20で、音源方向θ_{HMj}の音を分離する。音源分離は、通過帯域関数を利用して、音源方向θ_{HMj}のIPD及びIIDのそれぞれの上限値$\Delta\phi_h(f)$, $\Delta\rho_h(f)$及び下限値$\Delta\phi_l(f)$, $\Delta\rho_l(f)$を求め、前記式(16)の条件と、この上限値、下限値の条件とから、音源方向θ_{HMj}のスペクトルと推定されるサブバンド(選択スペクトル)を選択する。その後、選択サブバンドのスペクトルを逆FFTにより変換すれば、音声信号に変換できる。

[0095] 次に、特徴抽出部30は、音源分離部20が分離した選択スペクトルを、対数変換部31、メル周波数変換部32、コサイン変換部33によりMFCCに変換する。

[0096] 一方、音響モデル合成部40は、音響モデル記憶部49に記憶された方向依存音響モデル$H(\theta_n)$と、音源定位部10が定位した音源方向θ_{HMj}とから、音源方向θ_{HMj}に適切と考えられる音響モデルを合成する。

すなわち、音響モデル合成部40は、方向依存音響モデルH(θ_n)を、コサイン逆変換部41、線形変換部42、及び指数変換部43により、線形スペクトルに変換する。そして、パラメータ合成部44は、音源方向θ_{HMj}の重み$W_{n\theta HMj}$を音響モデル記憶部49から読み出し、これと方向依存音響モデルH(θ_n)との内積をとって、音源方向θ_{HMj}の音響モデルH(θ_{HMj})を合成する。そして、この線形スペクトルで表された音響モデルH(θ_{HMj})を、対数変換部45，メル周波数変換部46、及びコサイン変換部47によりMFCCで表した音響モデルH(θ_{HMj})に変換する。

[0097] 次に、音声認識部50は、音響モデル合成部40で合成された音響モデルH(θ_{HMj})を利用して、隠れマルコフモデルにより音声認識を行う。

[0098] このようにして、音声認識を行った結果の例が、表4である。

[0099] [表4]

	従来手法							本発明
音響モデルの方向	−90°	−60°	−30°	0	30°	60°	90°	40°
孤立単語認識率	20%	20%	38%	42%	60%	59%	50%	78%

[0100] 表4に示すように、方向依存音響モデルを−90° ー90° まで30° おきに用意して、各音響モデルで40° の方向から孤立単語を認識させたところ(従来手法)、最も認識率が高くても30° 方向の方向依存音響モデルを用いた60%であった。これに対し、本実施形態の手法を使用して40° 方向の音響モデルを合成して、これを用いて孤立単語を認識させたところ、78%の高い認識率を示した。このように、本実施形態の音声認識装置1によれば、任意の方向から音声が発せられた場合であっても、その方向に適した音響モデルをその都度合成するので、高い認識率を実現することができる。また、任意の方向の音声を認識できることから、移動している音源からの音声認識や、移動体(ロボットRB)自身が移動しているときにも、高い認識率での音声認識が可能である。

[0101] また、方向依存音響モデルを、断続的な数個、例えば音源方向にして60° ごとや30° ごとに記憶しておけば良く、音響モデルの学習に必要なコストを小さくすることができる。

さらに、合成した音響モデル一つについて音声認識を行えば良いため、複数方向

の音響モデルについて音声認識を試みる並列処理も不要であり、計算コストを小さくすることができる。そのため、実時間処理や、組み込み用途には好適である。

[0102] 以上、本発明の第1実施形態について説明したが、本発明は第1実施形態には限定されず、以下の実施形態のように変形して実施することが可能である。

[0103] [第2実施形態]

第2実施形態では、第1実施形態の音源定位部10に代えて、相互相関のピークを用いて音源方向を定位する音源定位部110を備える。なお、他の部分については第1実施形態と同様であるので説明を省略する。

《音源定位部110》

第2実施形態に係る音源定位部110は、図18に示すように、フレーム切り出し部111、相互相関計算部112、ピーク抽出部113、方向推定部114を有する。

[0104] 〈フレーム切り出し部111〉

フレーム切り出し部111は、左右のマイクM_R, M_Lに入力されたそれぞれの音響信号について、所定の時間長、例えば100msecで切り出す処理を行う。切り出し処理は、適当な時間間隔、例えば30msecごとに行われる。

[0105] 〈相互相関計算部112〉

相互相関計算部112は、フレーム切り出し部111が切り出した右マイクM_Rの音響信号と、左マイクM_Lの音響信号とで、次式(22)により相互相関を計算する

[数22]

$$CC(T) = \int_0^T x_L(t)x_R(t+T)dt \quad \cdots \quad (22)$$

但し、

CC(T):x_L(t)とx_R(t)の相互相関

T:フレーム長

x_L(t):フレーム長Tで切り出された、マイクLからの入力信号

x_R(t):フレーム長Tで切り出された、マイクRからの入力信号

[0106] 〈ピーク抽出部113〉

ピーク抽出部113は、得られた相互相関の結果からピークを抽出する。抽出するピ

ークの数は、音源の数が予め分かっている場合は、その数に対応したピークを大きいものから選択する。音源数が不明なときは、予め定めた閾値を超えたピークを全て抽出するか、あるいは予め定めた所定数のピークを大きいものから順に選択する。

[0107] 〈方向推定部114〉

音源方向θ_{HMj}は、得られたピークから、右マイクM_Rと左マイクM_Lに入力された音響信号の到達時間差Dに音速vを掛けて、図19に示す距離差dを計算し、さらに、次式により求める。

$\theta_{HMj} = \arcsin(d/2r)$

[0108] このような相互相関を用いた音源定位部110によっても、音源方向θ_{HMj}の方向が推定され、前記した音響モデル合成部40により、音源方向θ_{HMj}に適した音響モデルを合成することで、認識率の向上を図ることができる。

[0109] [第3実施形態]

第3実施形態では、第1実施形態に加えて、音源定位部音源が同一音源から来ていることを確認しながら音声認識を行う機能を追加している。なお、第1実施形態と同じ部分については、同じ符号を付して説明を省略する。

第3実施形態に係る音声認識装置100は、図20に示すように、第1実施形態の音声認識装置1に加え、音源定位部10が定位した音源方向を入力されて、音源を追跡し、同じ音源から音響が来続けているかを確認し、確認ができたなら、音源方向を音源分離部20へ出力するストリーム追跡部60を有している。

[0110] 図21に示すように、ストリーム追跡部60は、音源方向履歴記憶部61と、予測部62と、比較部63とを有する。

[0111] 音源方向履歴記憶部61は、図22に示すような、時間と、その時間において認識された音源の方向及び音源のピッチ(その音源の調波構造が持つ基本周波数f_0)とが関連づけて記憶されている。

[0112] 予測部62は、音源方向履歴記憶部61から、直前まで追跡していた音源の音源方向の履歴を読み出し、直前までの履歴からカルマンフィルタなどにより現時点t1での音源方向θ_{HMj}及び基本周波数f_0とからなるストリーム特徴ベクトル(θ_{HMj}, f_0)を予測し、比較部63へ出力する。

[0113] 比較部63は、音源定位部10から、音源定位部10で定位された現時点t1の各話者HMjの音源方向θ_{HMj}と、その音源の基本周波数f_0とが入力される。そして、予測部62から入力された予測したストリーム特徴ベクトル(θ_{HMj}, f_0)と、音源定位部10で定位された音源方向及びピッチから求まるストリーム特徴ベクトル(θ_{HMj}, f_0)を比較して、その差(距離)が予め定めた閾値よりも小さい場合に、音源方向θ_{HMj}を音源分離部に出力する。また、ストリーム特徴ベクトル(θ_{HMj}, f_0)を音源方向履歴記憶部61へ記憶させる。

前記した差(距離)が、予め定めた閾値よりも大きい場合には、定位した音源方向θ_{HMj}を音源分離部20へ出力しないので、音声認識は行われない。なお、音源方向θ_{HMj}とは別に、音源の追跡ができているか否かを示すデータを、比較部63から音源分離部20へ出力してもよい。

なお、基本周波数f_0を用いず、音源方向θ_{HMj}だけで予測してもよい。

[0114] このようなストリーム追跡部60を有する音声認識装置100によれば、音源定位部10で音源方向が定位され、ストリーム追跡部60へ音源方向とピッチが入力される。ストリーム追跡部60では、予測部62が、音源方向履歴記憶部61に記憶された音源方向の履歴を読み出して現時点t1でのストリーム特徴ベクトル(θ_{HMj}, f_0)を予測する。比較部63は、予測部62で予測されたストリーム特徴ベクトル(θ_{HMj}, f_0)と、音源定位部10から入力された値から求まるストリーム特徴ベクトル(θ_{HMj}, f_0)とを比較して、その差(距離)が所定の閾値より小さければ、音源方向を音源分離部20へ出力する。

音源分離部20は、音源定位部10から入力されたスペクトルのデータと、ストリーム追跡部60が出力した音源方向θ_{HMj}のデータに基づき、第1実施形態と同様にして音源を分離する。そして、以下、特徴抽出部30、音響モデル合成部40、音声認識部50でも、第1実施形態と同様にして、処理を行う。

[0115] このように、本実施形態の音声認識装置100は、音源が追跡できているか否かを確認した上で音声認識を行うので、音源が移動している場合にも、同じ音源が発し続けている音声を連続して認識するため、誤認識の可能性を低くすることができる。特に、複数の移動する音源があって、それらの音源が交差する場合などに好適である。

また、音源方向を記憶、予測していることから、その方向の所定範囲についてのみ

音源を探索すれば、処理を少なくすることができる。

[0116] 以上、本発明の実施形態について説明したが、本発明は、前記した実施形態には限定されず適宜変更して実施される。

例えば、音声認識装置1が、カメラと、公知の画像認識装置を有し、話者の顔を認識して、誰が話しているかを自己が有するデータベースから話者を特定する話者同定部を備え、前記方向依存音響モデルを話者ごとに有していれば、話者に適した音響モデルを合成することができるので、認識率をより高くする事ができる。あるいは、カメラを使わず、ベクトル量子化(VQ)を用いて、予め登録してある話者の音声をベクトル化したものと、音源分離部20で分離された音声をベクトル化したものとを比較し、最も距離の近い話者を結果として出力することで話者を同定してもよい。

請求の範囲

[1] 複数のマイクが検出した音響信号から、音声を認識して文字情報に変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、

前記複数のマイクが検出した1つ以上の音響信号に基づき、その音響信号に含まれる音声信号の特徴を抽出する特徴抽出部と、

断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えることを特徴とする音声認識装置。

[2] 複数のマイクが検出した音響信号から、特定の話者の音声を認識して文字情報に変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、

前記音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、

前記音源分離部が分離した音声信号の特徴を抽出する特徴抽出部と、

断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽

出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えることを特徴とする音声認識装置。

[3] 前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成されたことを特徴とする請求の範囲第1項または第2項に記載の音声認識装置。

[4] 前記音源定位部は、前記複数のマイクから検出された音響信号の音圧差と位相差を用いて前記特定の話者の音源方向を特定するために、前記マイクが設けられる部材の表面で散乱する音響信号を音源方向ごとにモデル化した散乱理論を用いることを特徴とする請求の範囲第1項から第3項のいずれか1項に記載の音声認識装置。

[5] 前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されたことを特徴とする請求の範囲第2項から第4項のいずれか1項に記載の音声認識装置。

[6] 前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、

前記線形和に使用する重みが、学習により決定されたことを特徴とする請求の範囲第1項から第5項のいずれか1項に記載の音声認識装置。

[7] 前記話者を特定する話者同定部をさらに備え、

前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、

前記音響モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、前記音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されたことを特徴とする請求項の範囲第1項から第6項のいずれか1項に記載の音声認識装置。

[8] 複数のマイクが検出した音響信号から、特定の話者の音声を認識して文字情報に

変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、

前記音源定位部が特定した音源方向を記憶して前記特定の話者の移動している方向を推定し、その推定された方向から、現在の話者の位置を推定するストリーム追跡部と、

前記ストリーム追跡部が推定した現在の話者の位置から定まる音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、

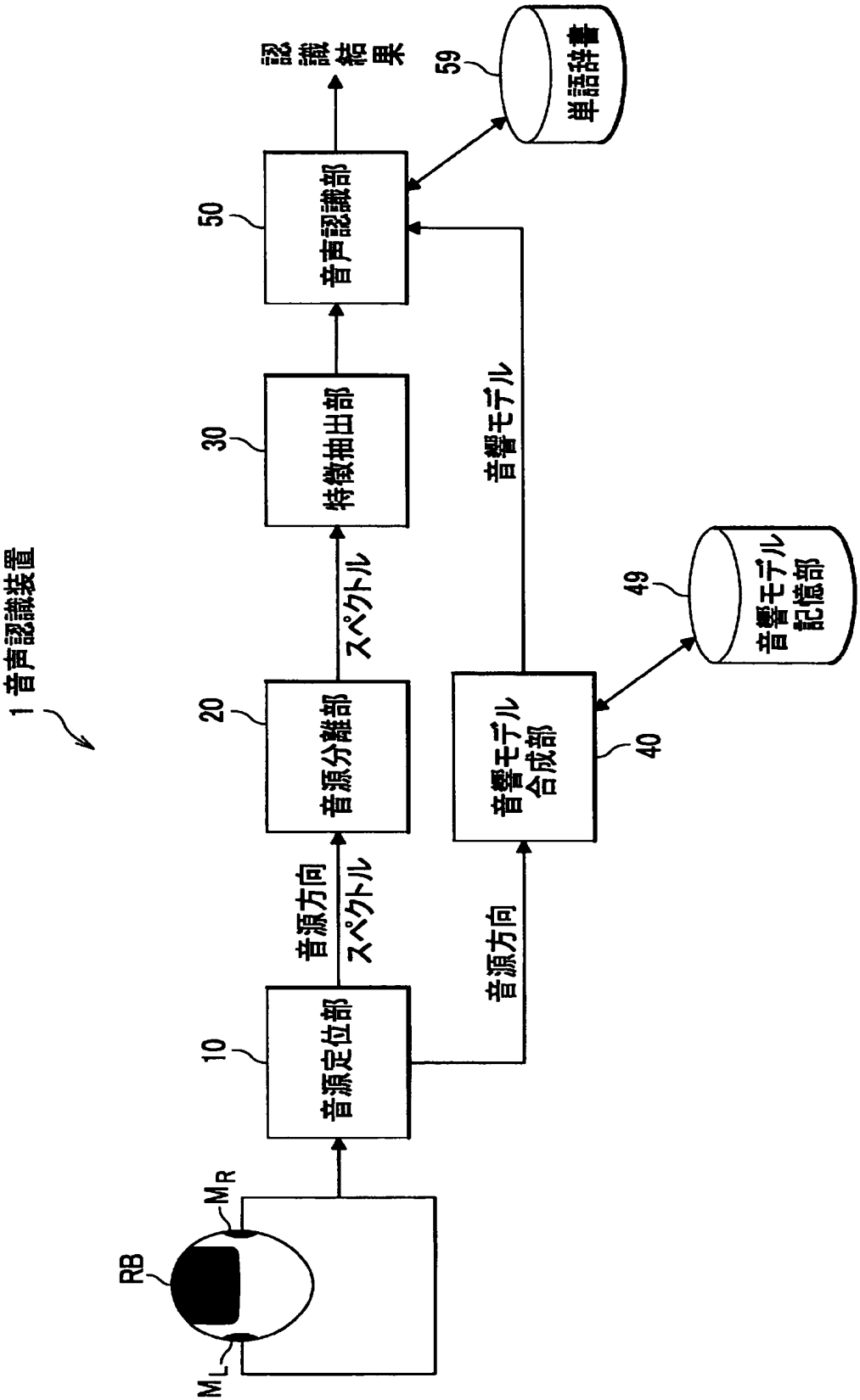
前記音源分離部が分離した音声信号の特徴を抽出する特徴抽出部と、

断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

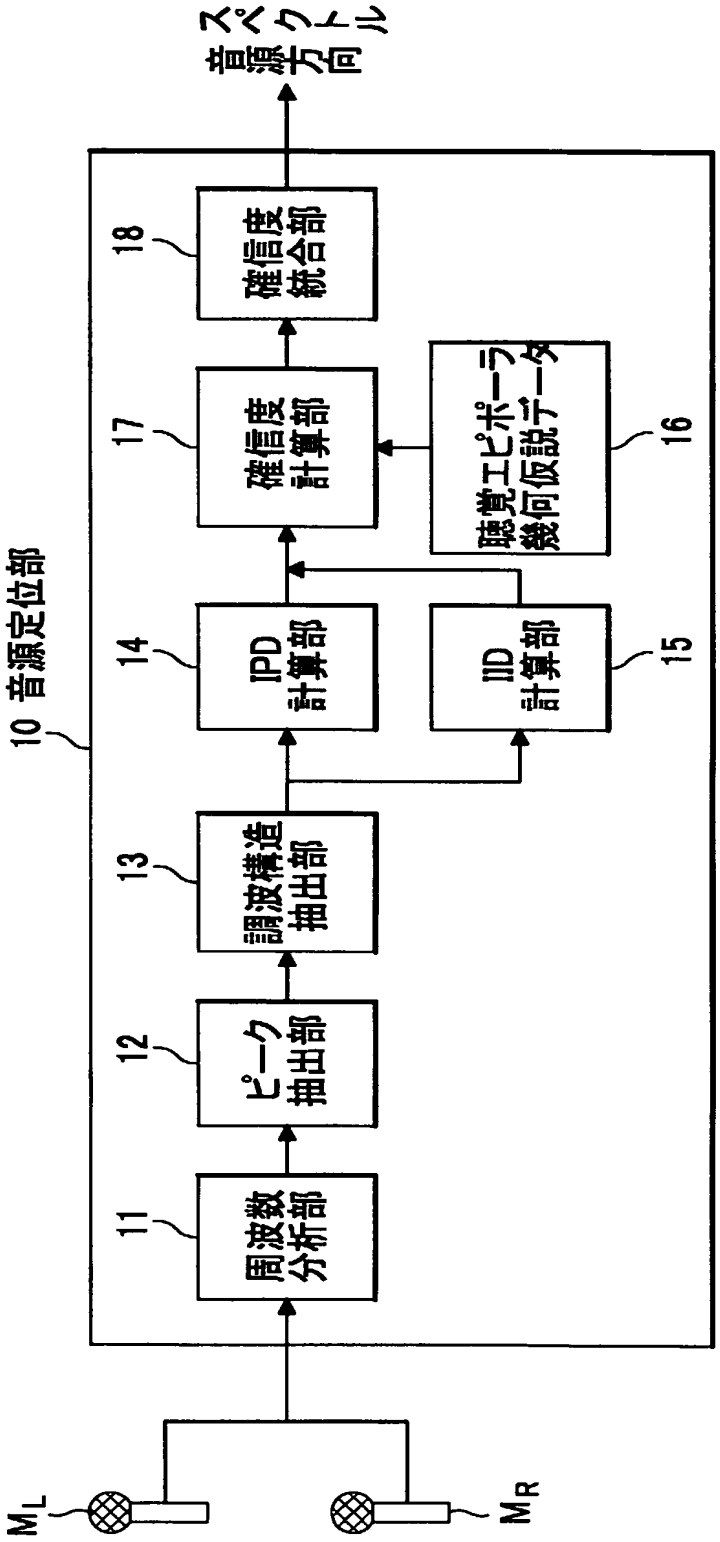
前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えることを特徴とする音声認識装置。

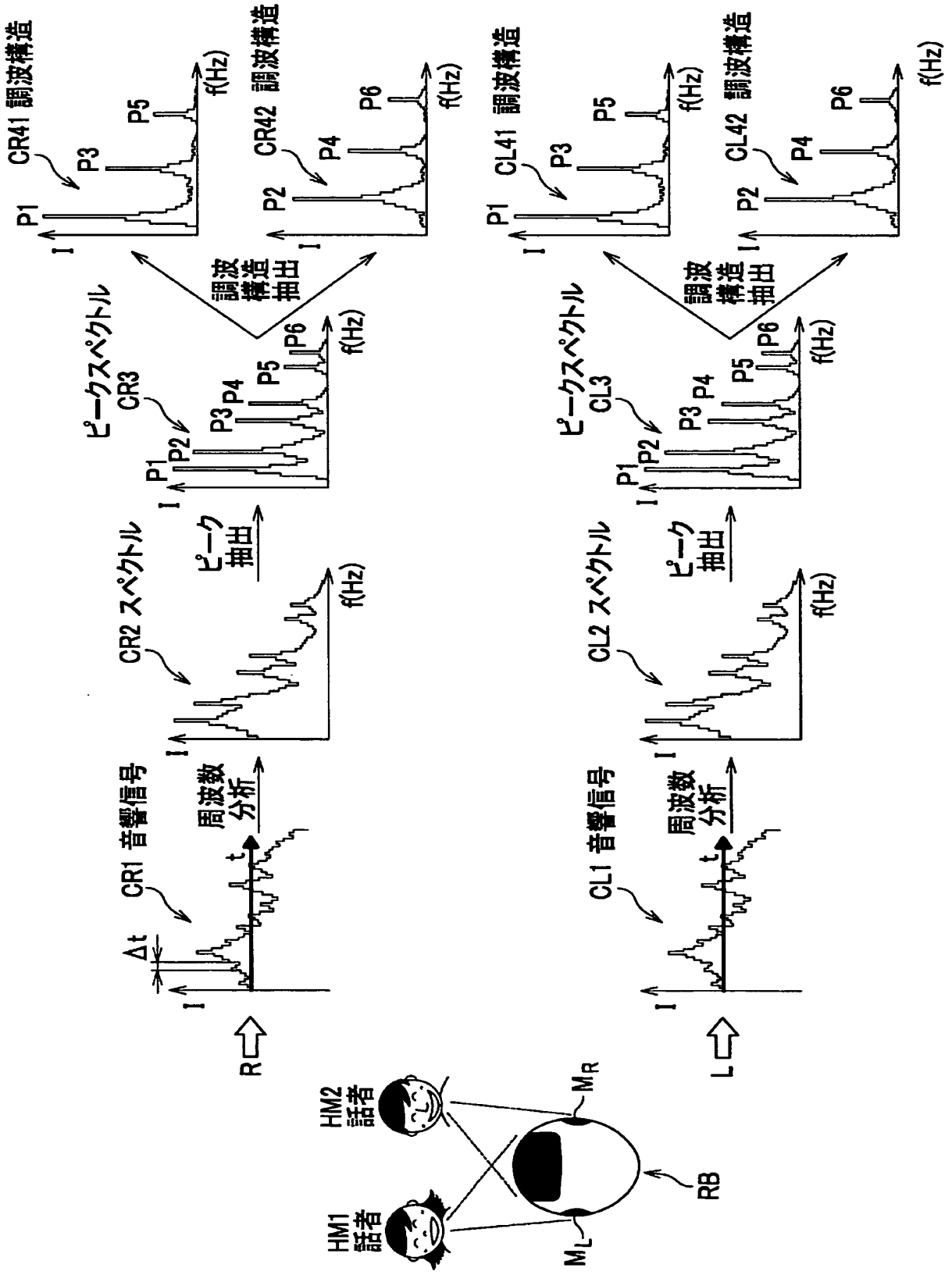
[図1]


1 音声認識装置
認識結果
59
単語辞書
50
音声認識部
30
特徴抽出部
スペクトル
音響モデル
49
音響モデル記憶部
20
音源分離部
音響モデル合成部
40
音源方向
スペクトル
音源方向
10
音源定位部
RB
M_R
M_L

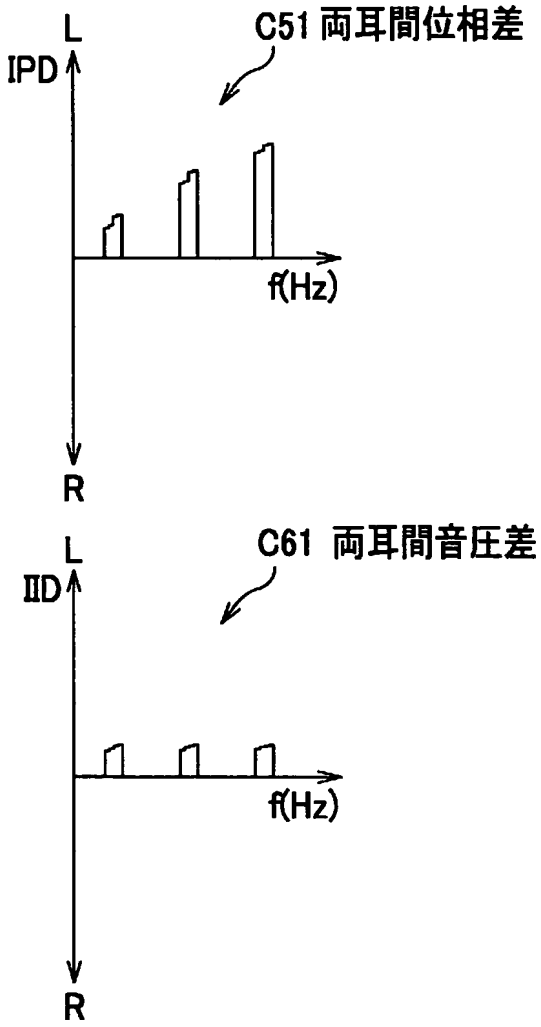
[図2]


スペクトル
音源方向
10 音源定位部
18
確信度統合部
17
確信度計算部
16
聴覚エピポーラ幾何仮説データ
14
IPD計算部
15
IID計算部
13
調波構造抽出部
12
ピーク抽出部
11
周波数分析部
M_L
M_R

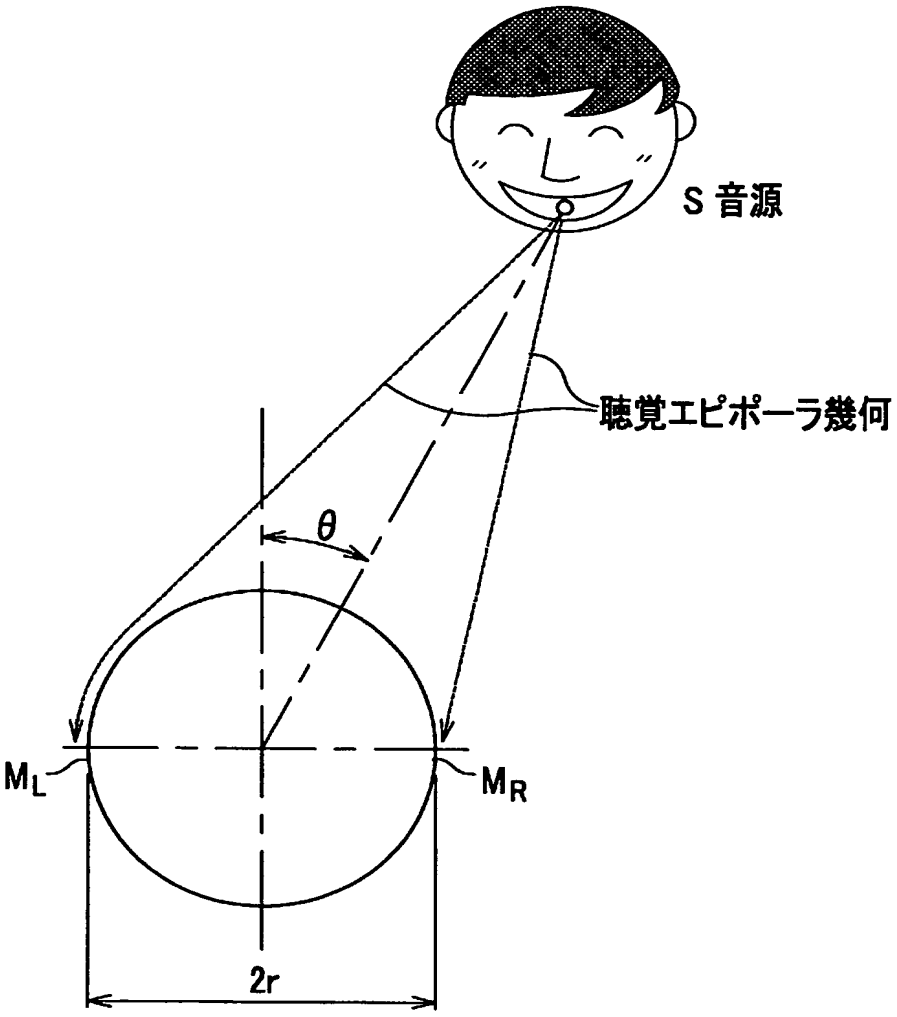
[図3]


CR41 調波構造
CR42 調波構造
CL41 調波構造
CL42 調波構造
P1
P2
P3
P4
P5
P6
f(Hz)
I
調波構造抽出
ピークスペクトル
CR3
CL3
ピーク抽出
CR2 スペクトル
CL2 スペクトル
CR1 音響信号
CL1 音響信号
周波数分析
Δt
t
R
L
HM2 話者
HM1 話者
M_R
M_L
RB

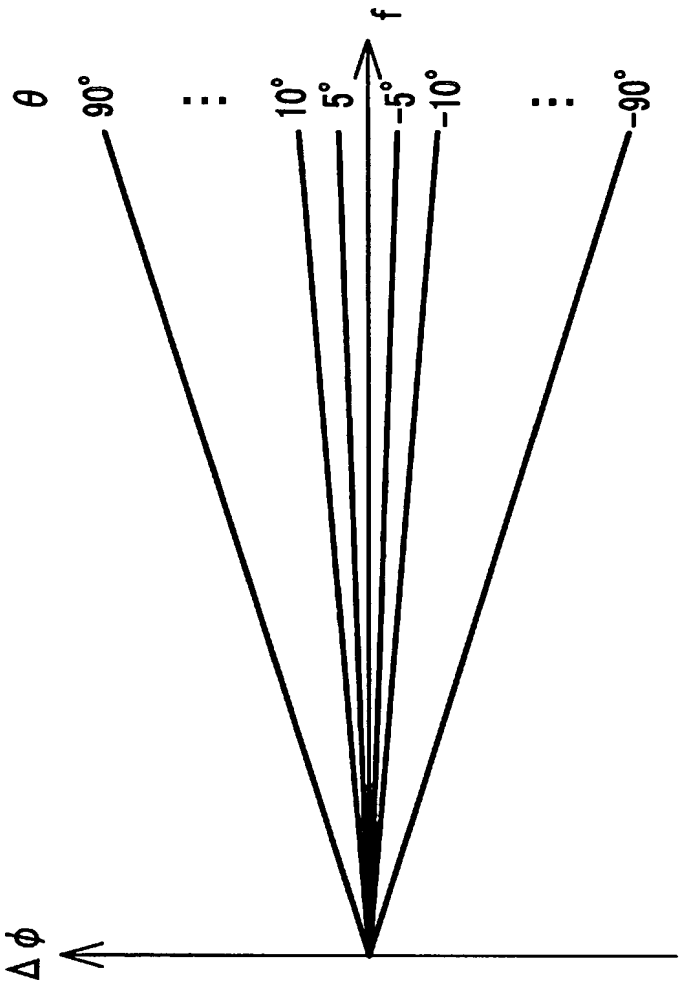
[図4]



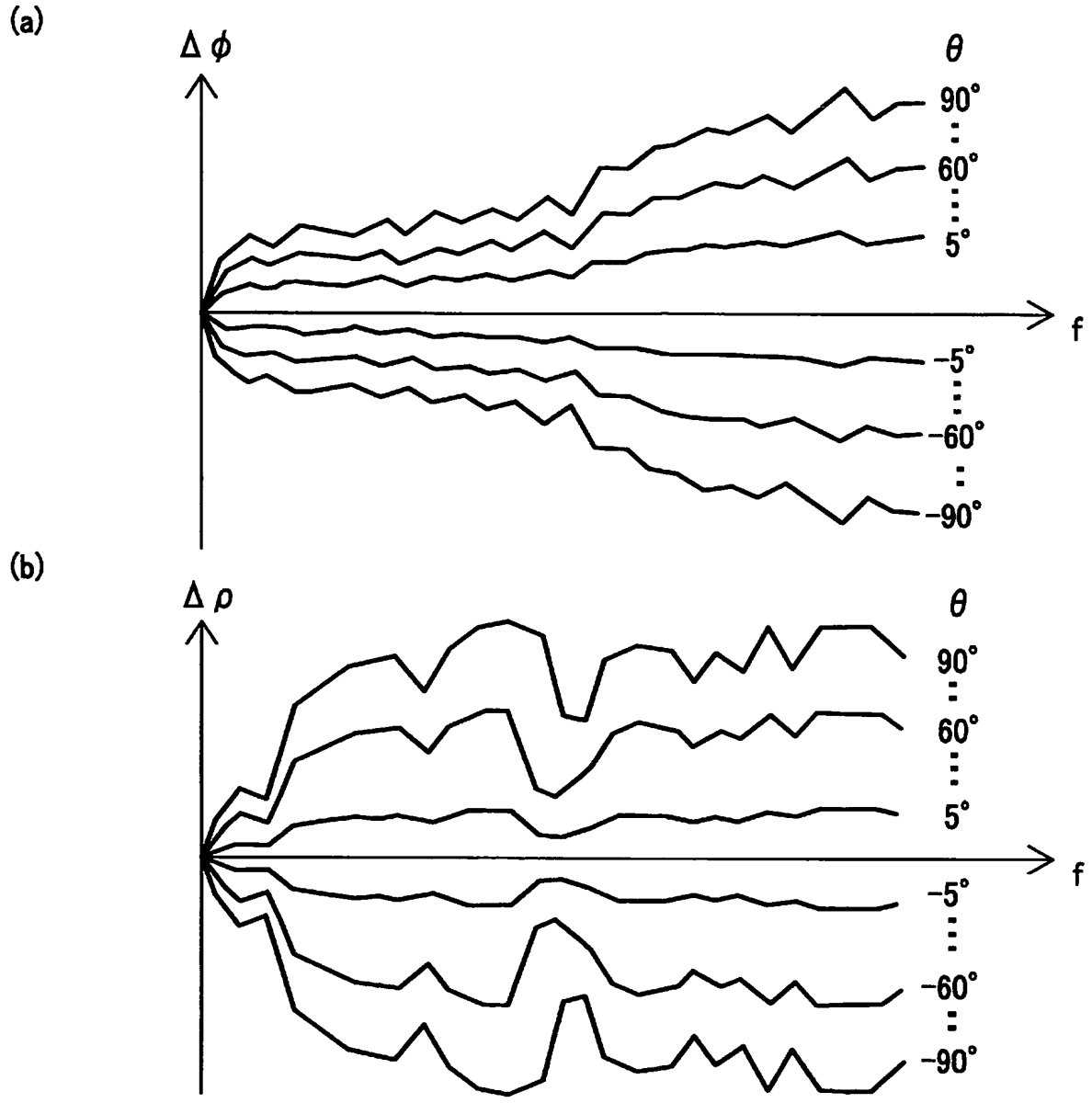
[図5]



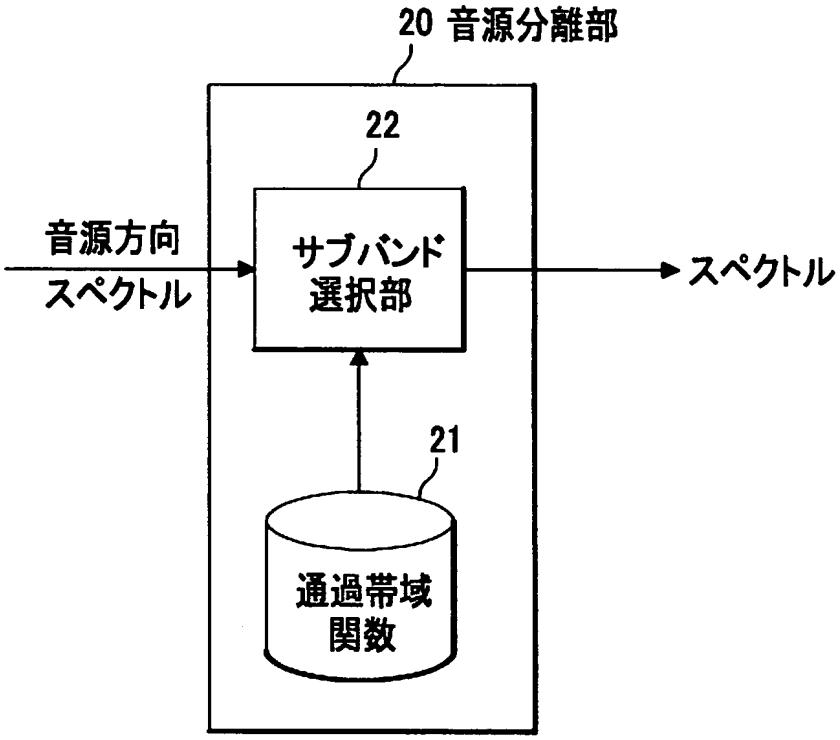
[図6]



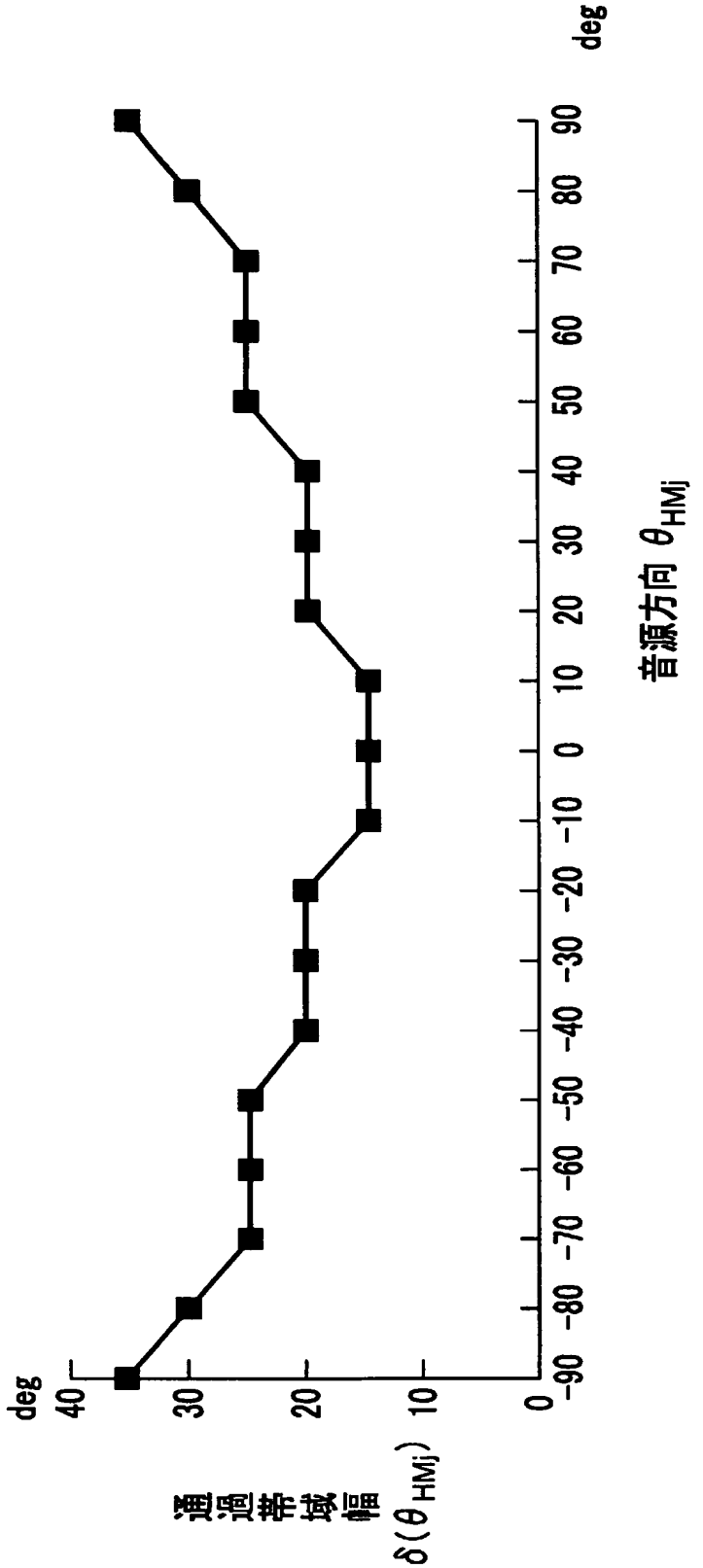
[図7]



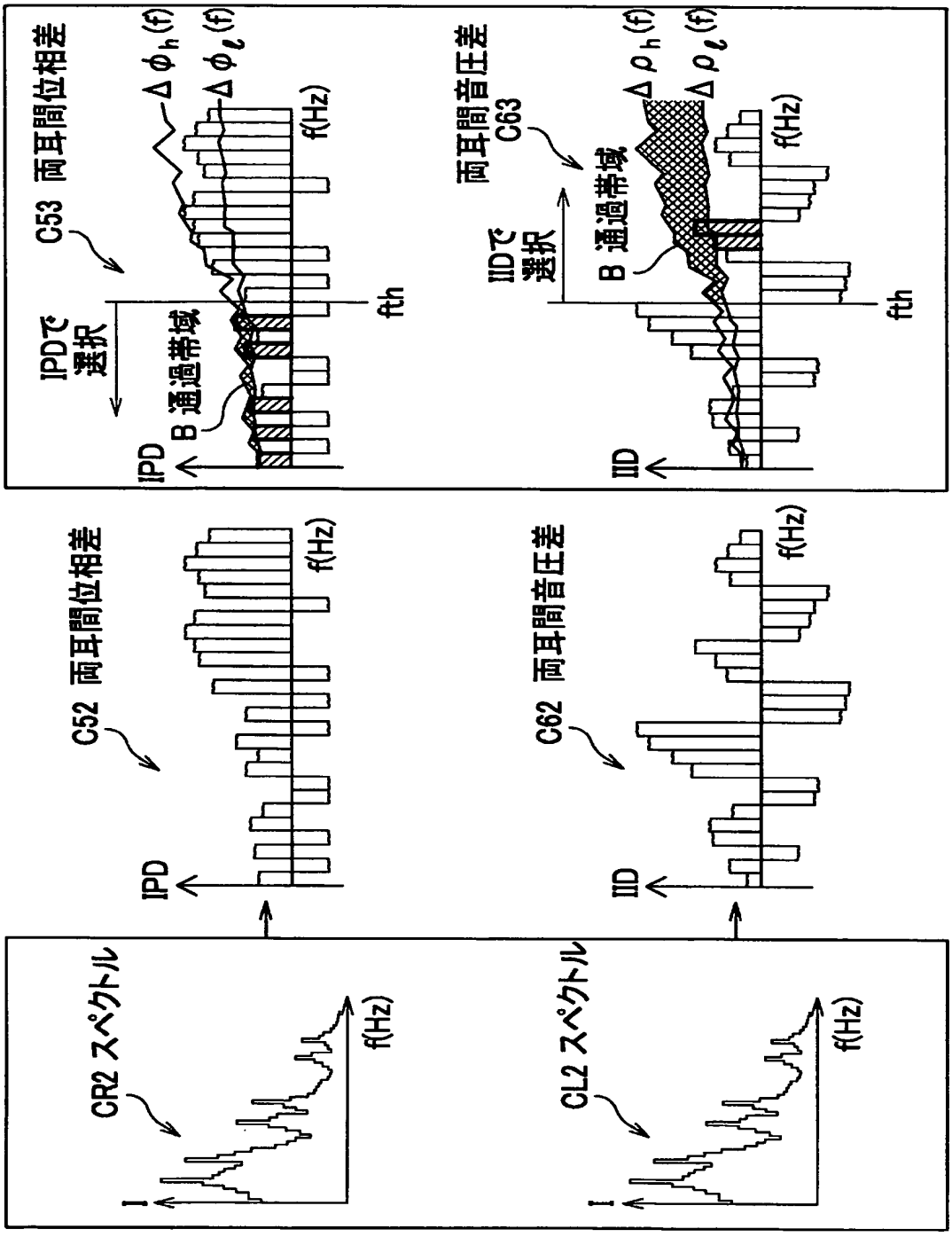
[図8]



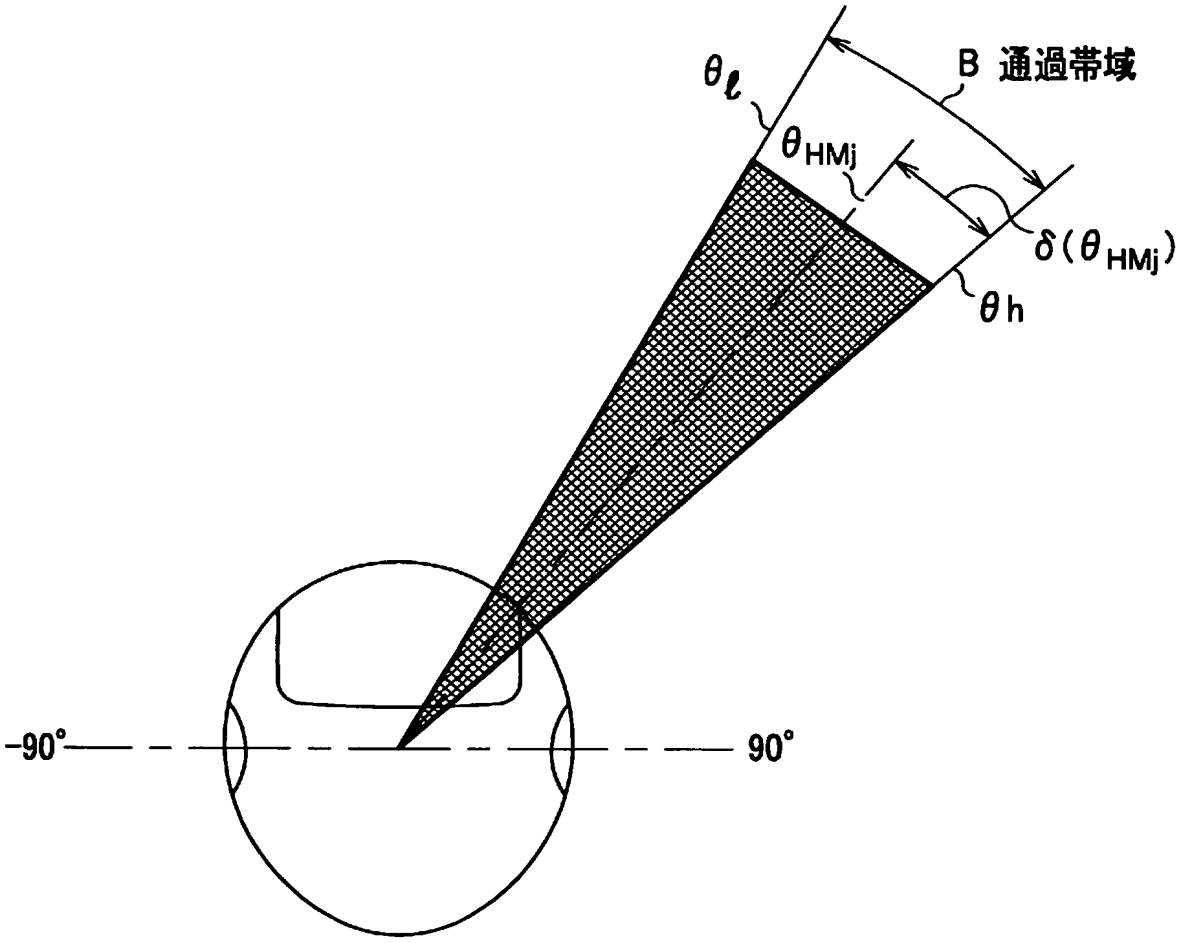
[図9]


deg
40
30
20
10
0
-90 -80 -70 -60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90
deg
音源方向 θ_{HMj}
聞き漏らし量
$\delta(\theta_{HMj})$

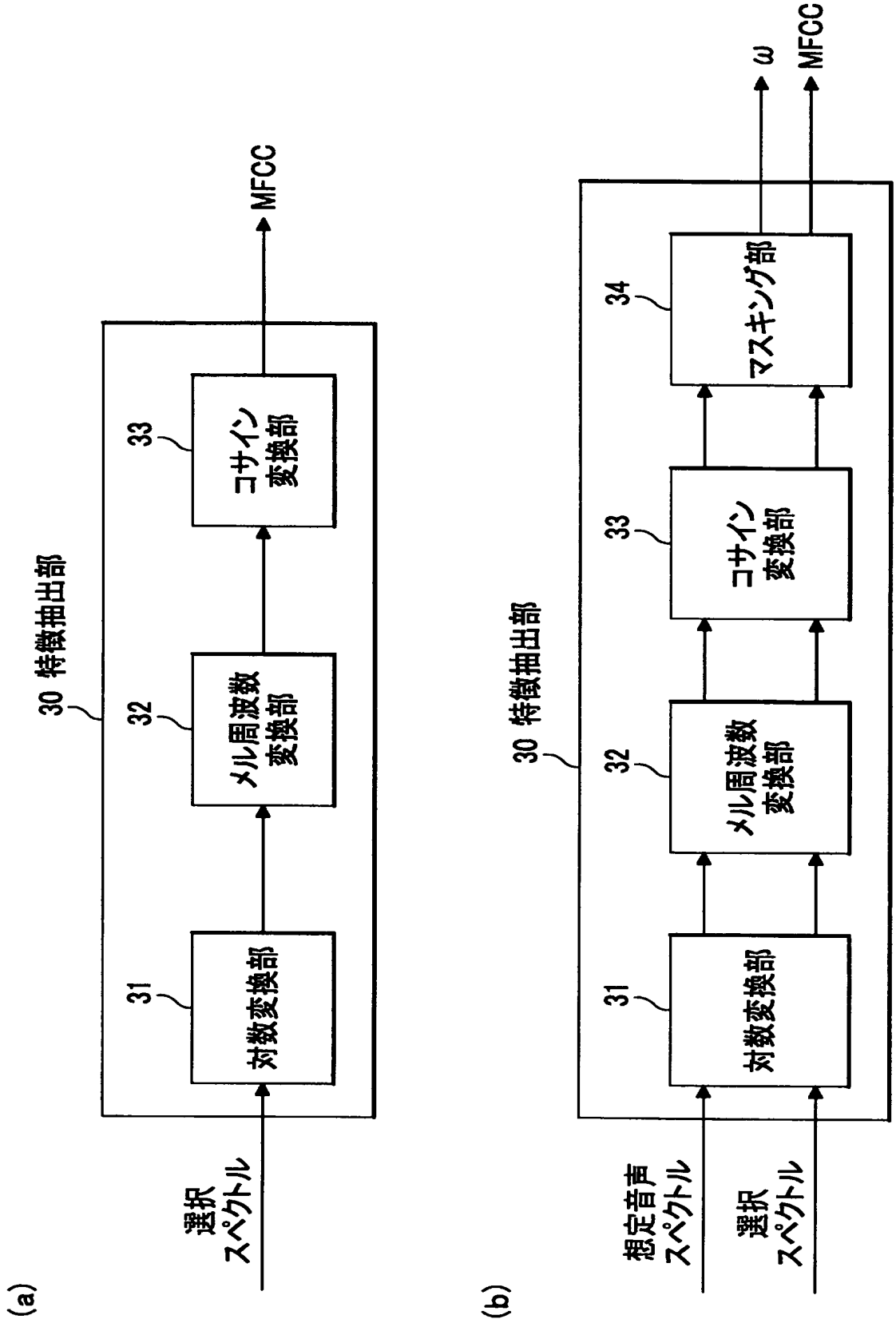
[図10]


C53 両耳間位相差
Δφh(f)
Δφℓ(f)
f(Hz)
IPDで選択
B通過帯域
fth
IPD
C63 両耳間音圧差
Δρh(f)
Δρℓ(f)
IIDで選択
IID
C52 両耳間位相差
C62 両耳間音圧差
CR2 スペクトル
CL2 スペクトル
I

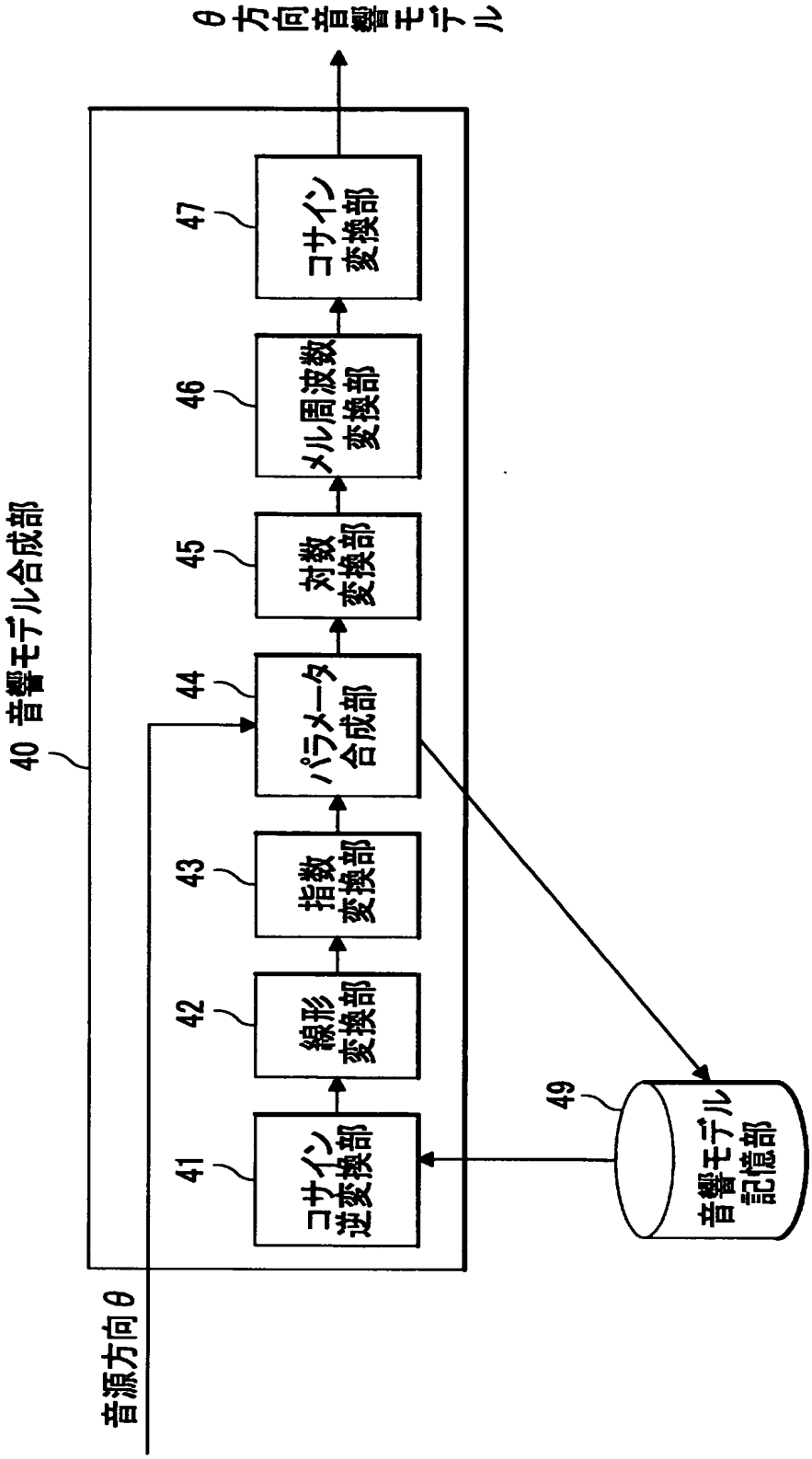
[図11]



[図12]


(a)
選択スペクトル
30 特徴抽出部
31 対数変換部
32 メル周波数変換部
33 コサイン変換部
MFCC
(b)
想定音声スペクトル
選択スペクトル
30 特徴抽出部
31 対数変換部
32 メル周波数変換部
33 コサイン変換部
34 マスキング部
MFCC
ω

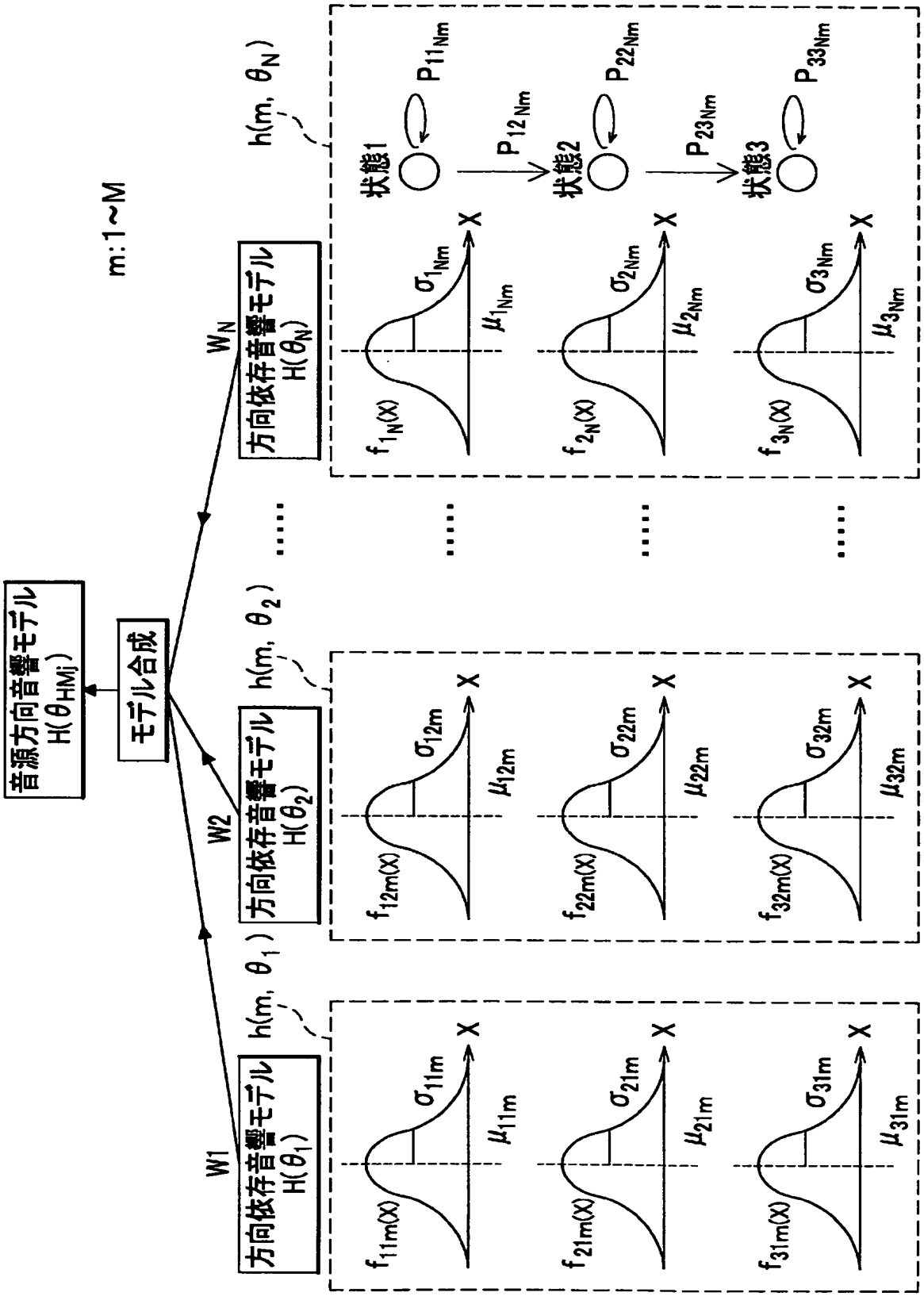
[図13]


θ方向音響モデル
40 音響モデル合成部
47
コサイン変換部
46
メル周波数変換部
45
対数変換部
44
パラメータ合成部
43
指数変換部
42
線形変換部
41
コサイン逆変換部
49
音響モデル記憶部
音源方向θ

[図14]

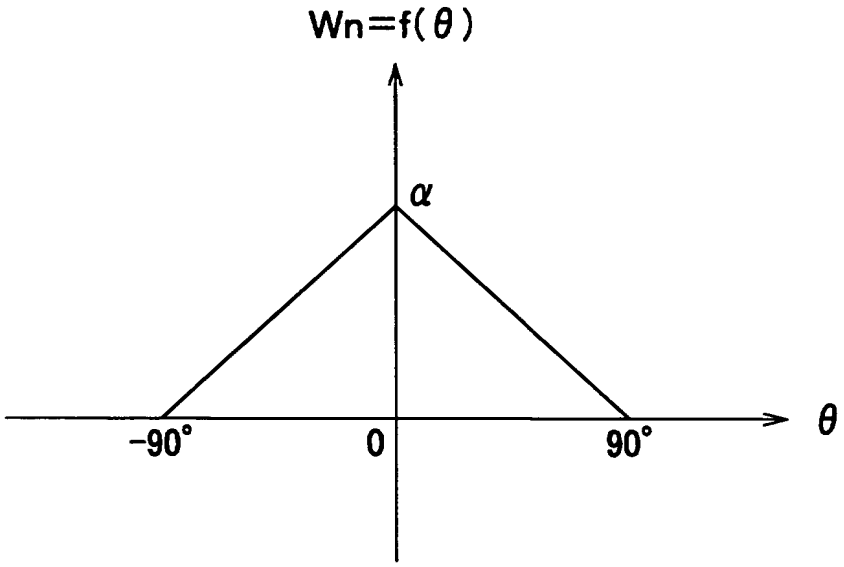
認識単位	サブモデル
/a/	h(/a/, θ n)
/b/	h(/b/, θ n)
⋮	⋮
m	h(m, θ n)
⋮	⋮

[図15]

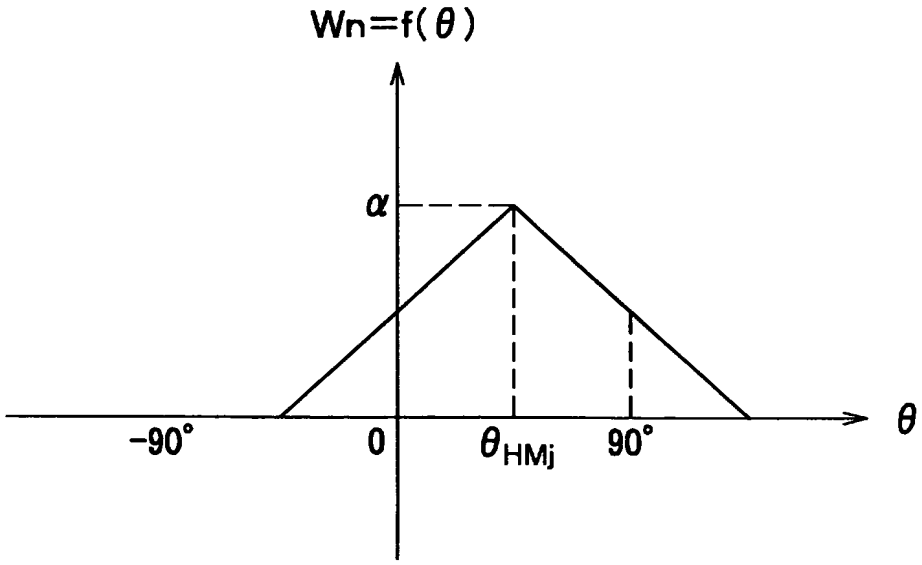

音源方向音響モデル
H(θHMj)
モデル合成
m:1~M
W1
W2
WN
方向依存音響モデル
H(θ1)
方向依存音響モデル
H(θ2)
方向依存音響モデル
H(θN)
h(m, θ1)
h(m, θ2)
h(m, θN)
状態1
状態2
状態3
P11Nm
P12Nm
P22Nm
P23Nm
P33Nm
f11m(X)
f21m(X)
f31m(X)
f12m(X)
f22m(X)
f32m(X)
f1N(X)
f2N(X)
f3N(X)
σ11m
σ21m
σ31m
σ12m
σ22m
σ32m
σ1Nm
σ2Nm
σ3Nm
μ11m
μ21m
μ31m
μ12m
μ22m
μ32m
μ1Nm
μ2Nm
μ3Nm
X

[図16]

(a)



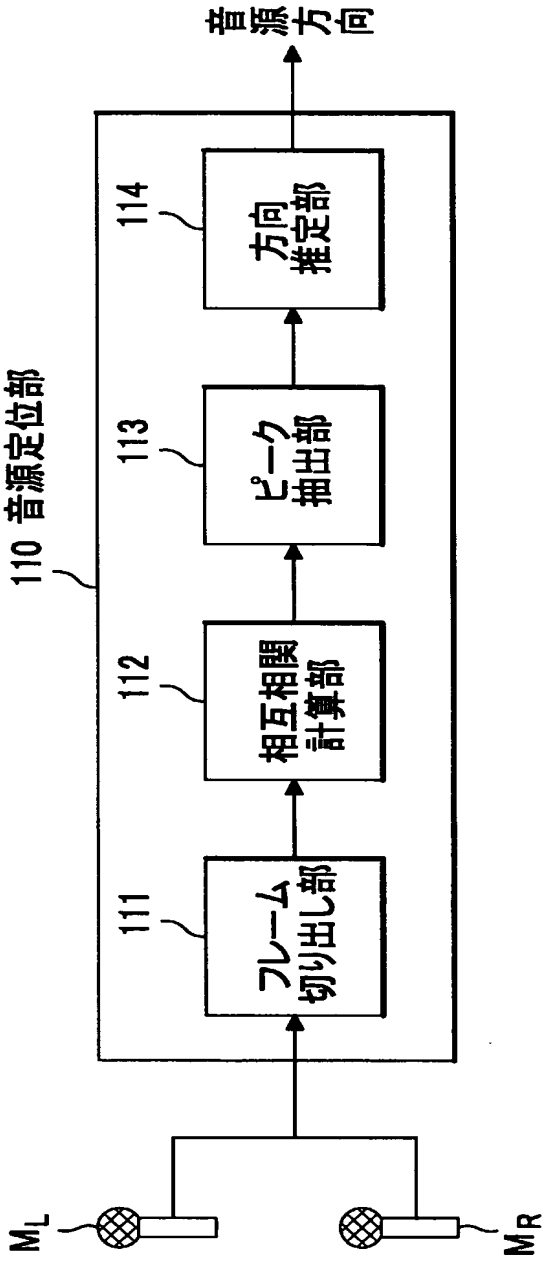
(b)



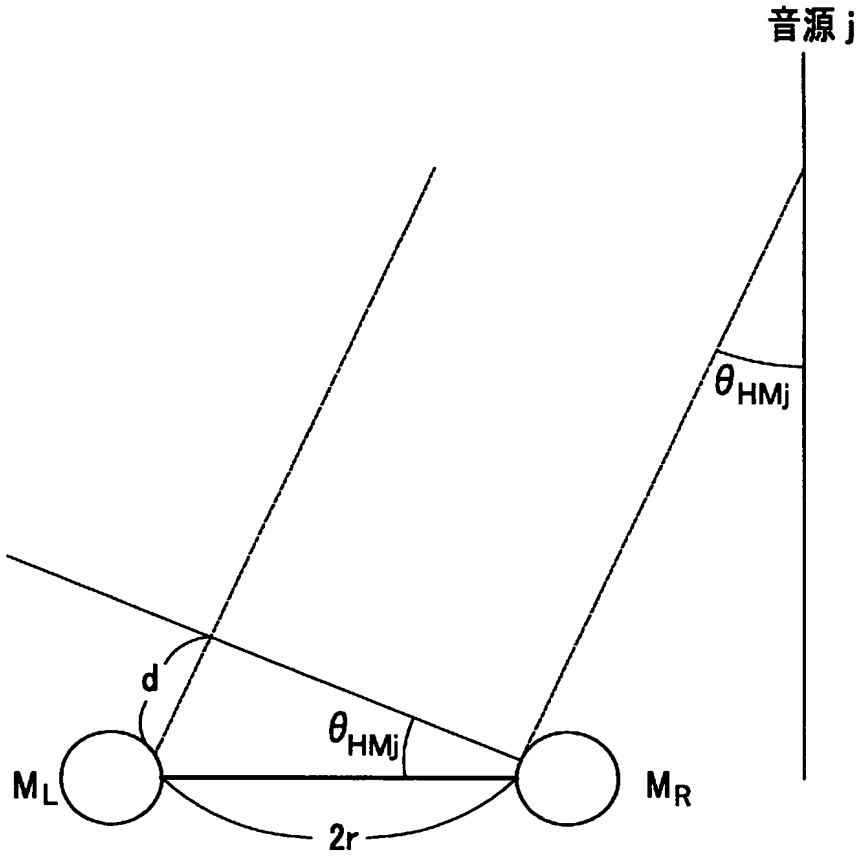
[図17]

認識単位	m	m'	m''
$H(\theta_{HMj})$	/x/	/y/	/z/
$H(\theta_{-90})$	/x/	/y/	m''
⋮	⋮	⋮	⋮
$H(\theta n)$	m	/y/	/z/
⋮	⋮	⋮	⋮
$H(\theta_{90})$	m	/y/	m''

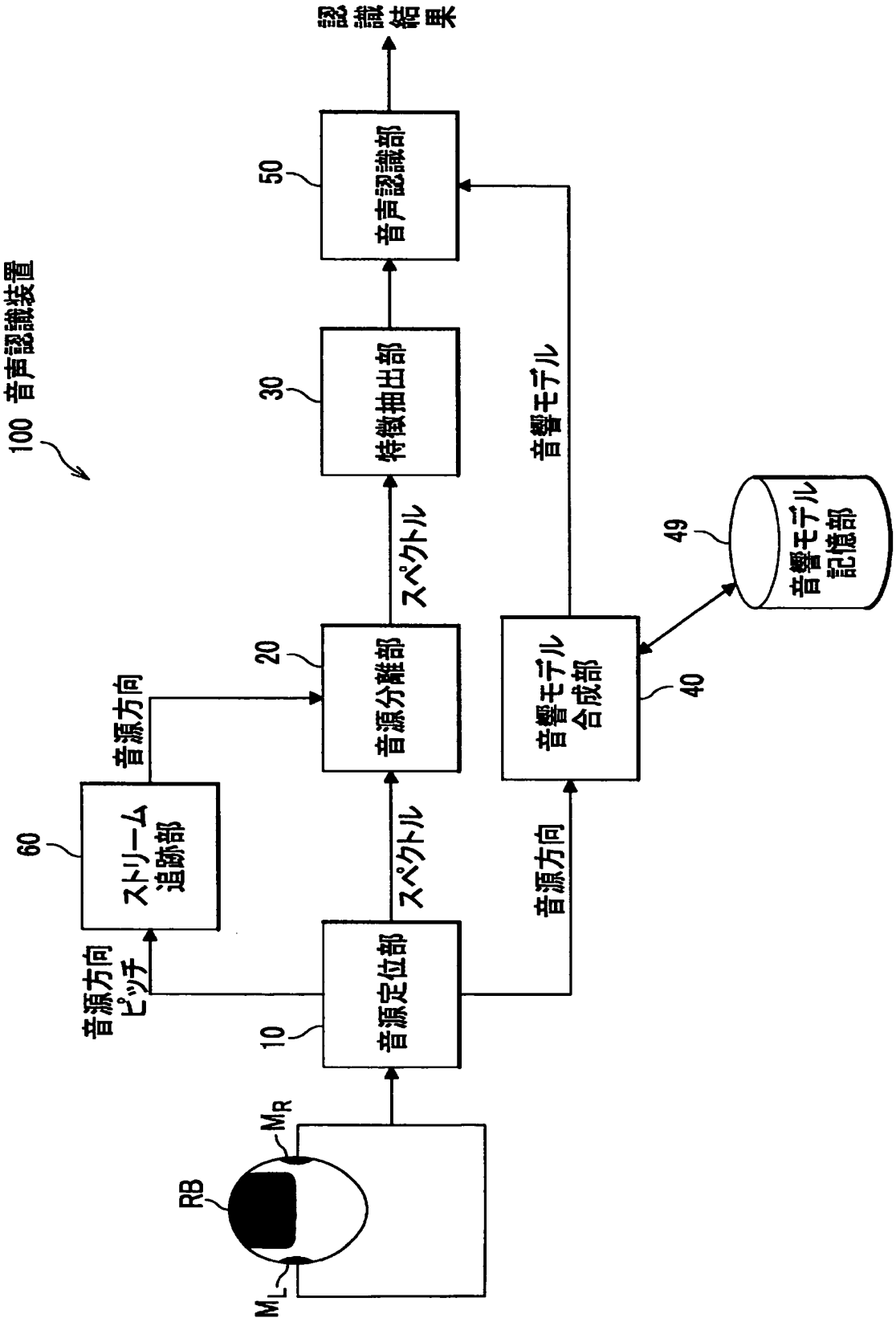
[図18]


音源方向
110 音源定位部
111
フレーム切り出し部
112
相互相関計算部
113
ピーク抽出部
114
方向推定部
M_L
M_R

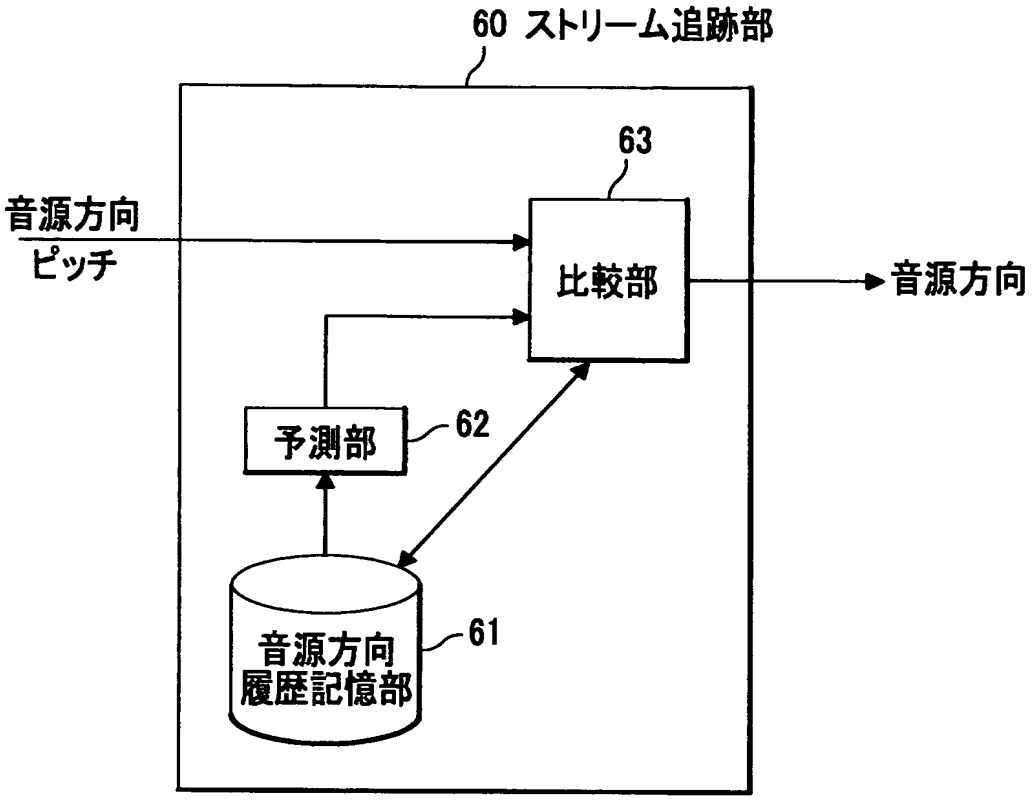
[図19]



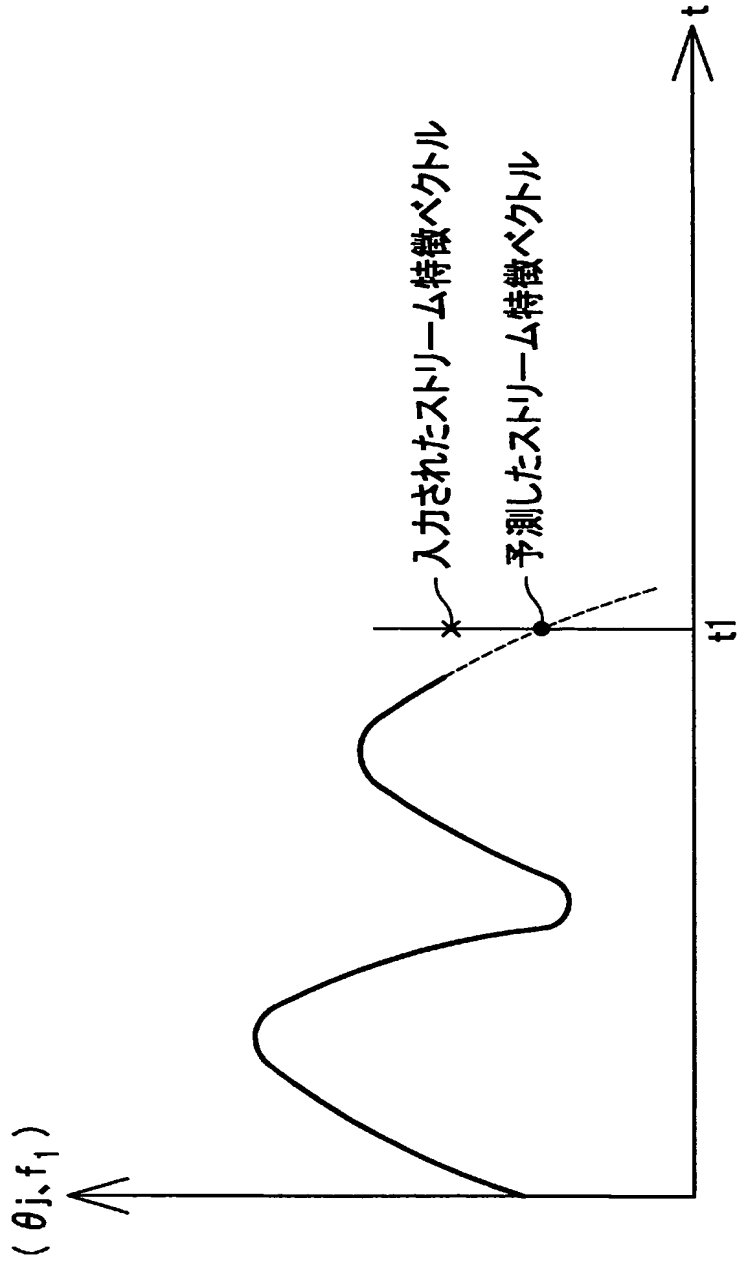
[図20]


認識結果
100 音声認識装置
50 音声認識部
30 特徴抽出部
スペクトル
音響モデル
49 音響モデル記憶部
20 音源分離部
音源方向
音響モデル合成部
40
60 ストリーム追跡部
音源方向
ピッチ
スペクトル
音源方向
10 音源定位部
RB
M_R
M_L

[図21]



[図22]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/016883

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G10L15/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G10L15/06, 15/20, 15/28, 21/02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2005
Kokai Jitsuyo Shinan Koho	1971-2005	Toroku Jitsuyo Shinan Koho	1994-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus FILE(JOIS)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2000-66698 A (Nippon Telegraph And Telephone Corp.), 03 March, 2000 (03.03.00), Full text; all drawings (Family: none)	1-8
Y	Kazuhiro NAKADAI, Daisuke MATSUURA, Hiroshi OKUNO, Koji TSUJINO, "Kaisoteki na Shichokaku togo to Sanran Riron o Riyo shita Robot ni yoru Sanwasha Doji Hatsuwa Ninshiki no Kojo", The Robotics Society of Japan Gakujutsu Koenkai Yokoshu, 20 September, 2003 (20.09.03), 2K14	1-8

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 02 February, 2005 (02.02.05)	Date of mailing of the international search report 15 February, 2005 (15.02.05)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/016883

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2002-264051 A (Japan Science and Technology Corp.), 18 September, 2002 (18.09.02), Full text; all drawings & JP 2000-264052 A & JP 2000-264053 A & JP 2000-264058 A & WO 2002/072317 A1 & EP 001375084 A1 & US 2004/0104702 A1	3-7
Y	Kazuhiro NAKADAI, Hiroshi OKUNO, Hiroaki KITANO, "Active Audition ni yoru Fukusu Ongen no Teii Bunri Ninshiki", The Japanese Society for Artificial Intelligence Kenkyukai Shiryo SIG-Challenge-0216-5, 22 November, 2002 (22.11.02), pages 25 to 32	5-8
Y	JP 11-143486 A (Fuji Xerox Co., Ltd.), 28 May, 1999 (28.05.99), Full text; all drawings (Family: none)	6,7
Y	JP 8-44387 A (Equos Research Co., Ltd.), 16 February, 1996 (16.02.96), Full text; all drawings (Family: none)	7
P,Y	Kazuhiro NAKADAI, Hiroshi OKUNO, Koji TSUJINO, "Robot o Taisho to shita Sanran Riron ni yoru Sanwasha Doji Hatsuwa no Teii Bunri Ninshiki no Kojo", The Japanese Society for Artificial Intelligence, Kenkyukai Shiryo SIG-Challenge-0318-6, 13 November, 2003 (13.11.03), pages 33 to 38	1-8
A	JP 2002-41079 A (Sharp Corp.), 08 February, 2002 (08.02.02), Full text; all drawings (Family: none)	1-8
A	JP 3-274593 A (Ricoh Co., Ltd.), 05 December, 1991 (05.12.91), Full text; all drawings (Family: none)	1
A	Kazuhiro NAKAGAI, Hiroshi OKUNO, Hiroaki KITANO, "Robot Recognizes Three Simultaneous Speech By Active Audition", Proc. of the 2003 IEEE, 14 September, 2003 (14.09.03), pages 398 to 405	1-8

国際調査報告

国際出願番号　PCT／JP2004／016883

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷ G10L15／06

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷ G10L15／06, 15／20, 15／28, 21／02

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報　1922－1996
日本国公開実用新案公報　1971－2005
日本国実用新案登録公報　1996－2005
日本国登録実用新案公報　1994－2005

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JTSPlusファイル（JOIS）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	JP 2000-66698 A （日本電信電話株式会社） 2000.03.03, 全文，全図 （ファミリーなし）	1-8
Y	中臺一博，松浦大輔，奥乃博，辻野広司，"階層的な視聴覚統合と散乱理論を利用したロボットによる三話者同時発話認識の向上"，日本ロボット学会学術講演会予稿集，2003.09.20，2K14	1-8

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
02.02.2005	15.2.2005

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）		
日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	山下　剛史	5C	3352
	電話番号　03－3581－1101　内線　3541		

様式PCT／ISA／210（第2ページ）（2004年1月）

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	JP 2002-264051 A (科学技術振興事業団) 2002.09.18, 全文，全図 & JP 2000-264052 A & JP 2000-264053 A & JP 2000-264058 A & WO 2002/072317 A1 & EP 001375084 A1 & US 2004/0104702 A1	3-7
Y	中臺一博，奥乃博，北野宏明， "アクティブオーディションによる複数音源の定位・分離・認識"， 人工知能学会研究会資料 SIG-Challenge-0216-5，2002.11.22， p.25-32	5-8
Y	JP 11-143486 A (富士ゼロックス株式会社) 1999.05.28, 全文，全図 (ファミリーなし)	6,7
Y	JP 8-44387 A (株式会社エクォス・リサーチ) 1996.02.16, 全文，全図 (ファミリーなし)	7
P，Y	中臺一博，奥乃博，辻野広司， "ロボットを対象とした散乱理論による三話者同時発話の定位・分離・認識の向上"， 人工知能学会研究会資料 SIG-Challenge-0318-6，2003.11.13， p.33-38	1-8
A	JP 2002-41079 A (シャープ株式会社) 2002.02.08, 全文，全図 (ファミリーなし)	1-8
A	JP 3-274593 A (株式会社リコー) 1991.12.05, 全文，全図 (ファミリーなし)	1
A	Kazuhiro NAKAGAI, Hirosh OKUNO, Hiroaki KITANO, "Robot Recognizes Three Simultaneous Speech By Active Audition", Proc. of the 2003 IEEE, 2003.09.14, p.398-405	1-8